Lustre / Lustre-HSM @ CINES

Olivier Rouchon – 9èmes journées meso-centres

12 Octobre 2016

Centre Informatique National de l'Enseignement Supérieur

Fournit à la communauté ESR des ressources, services et expertises informatiques exceptionnelles

3 activités principales:

- ✓ Calcul intensif
- ✓ Archivage pérenne
- ✓ Hébergement

✓Basé à Montpellier, France – approx. 60 personnes (ingénieurs, techniciens, administratif)

✓Créé en 1999, initialement **CNUSC** (Centre National Universitaire Sud de Calcul) – créé en 1980

✓Sous la tutelle directe du Ministère de l'Enseignement Supérieur et de la Recherche (MESR)

Infrastructures sécurisées

- 5 salles machines : 1 400 m²
- Locaux techniques : 2000 m²
- 2 lignes ERDF pour un total de 12,5 MW
- Données en double alimentation ondulée et sécurisée par groupe électrogène
- Copies et sauvegardes dans des salles distinctes + copie à distance

Ressources

- Un supercalculateur de niveau mondial
- Capacités de stockage de plusieurs PetaOctets
- Des accès réseau performants
- Des équipes d'experts

L'environnement de calcul

RENATER

Ressources pre/post traitement

CRISTAL : Bullx s6030 - 13.1 Tflops

2 nœuds s6030
- 32 cœurs : Intel Nehalem X7560 @ 2.27 GHz
- 256 GB RAM (nœud1) et 128 GB RAM (nœud2)
- 2 GPU nVIDIA Quadro FX5800

1 nœud R428
- 32 cœurs : Intel Sandy Bridge E5-4620 @ 2.20 GHz
- 1 TB RAM
- 4 GPU nVIDIA Quadro FX7000

8 nœuds R425
- 16 cœurs : Intel Sandy Bridge E5-2660 @ 2.20 GHz
- 128 GB / nœud
- 2 GPU nVIDIA Quadro FX6000

Réseau 10 Gb

Ressources HPC

OCCIGEN : Bullx DLC - 2.1 Pflops

2106 nœuds B720
- 50544 cœurs : Intel Haswell E5-2690 v3 @ 2.6 GHz
- Répartition 2,6 et 5,3 GB/cœur sur 50% des nœuds de calcul
- Infiniband FDR
- /scratch : Lustre, 5,2 PB @ 100 GB/s
- /home : Panasas, 400 TB @ 5 GB/s

CINES Centre Informatique National de l'Enseignement Supérieur

Passerelle Infiniband FDR

/store : Lustre : 2 PB @ 50 GB/s
NFS : 2 PB @ 2x10 GB/s

Librairie : 2 x IBM TS3500 ~2,5 PB
- 2 x 2000 cartouches
- 7 lecteurs Jaguar 3
- 7 lecteurs LTO 4

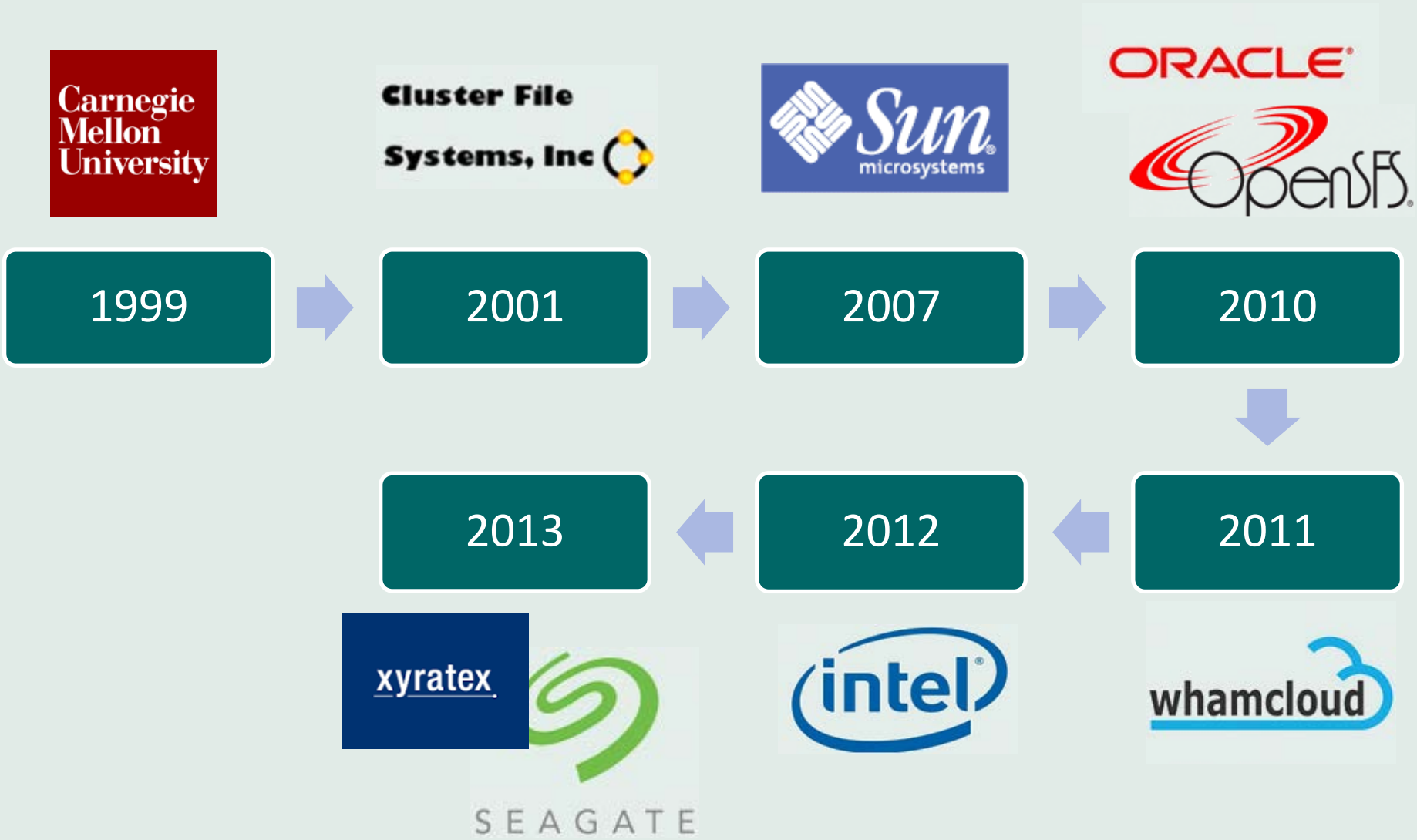
Ressources stockage / données

Le système de fichiers distribué libre Lustre

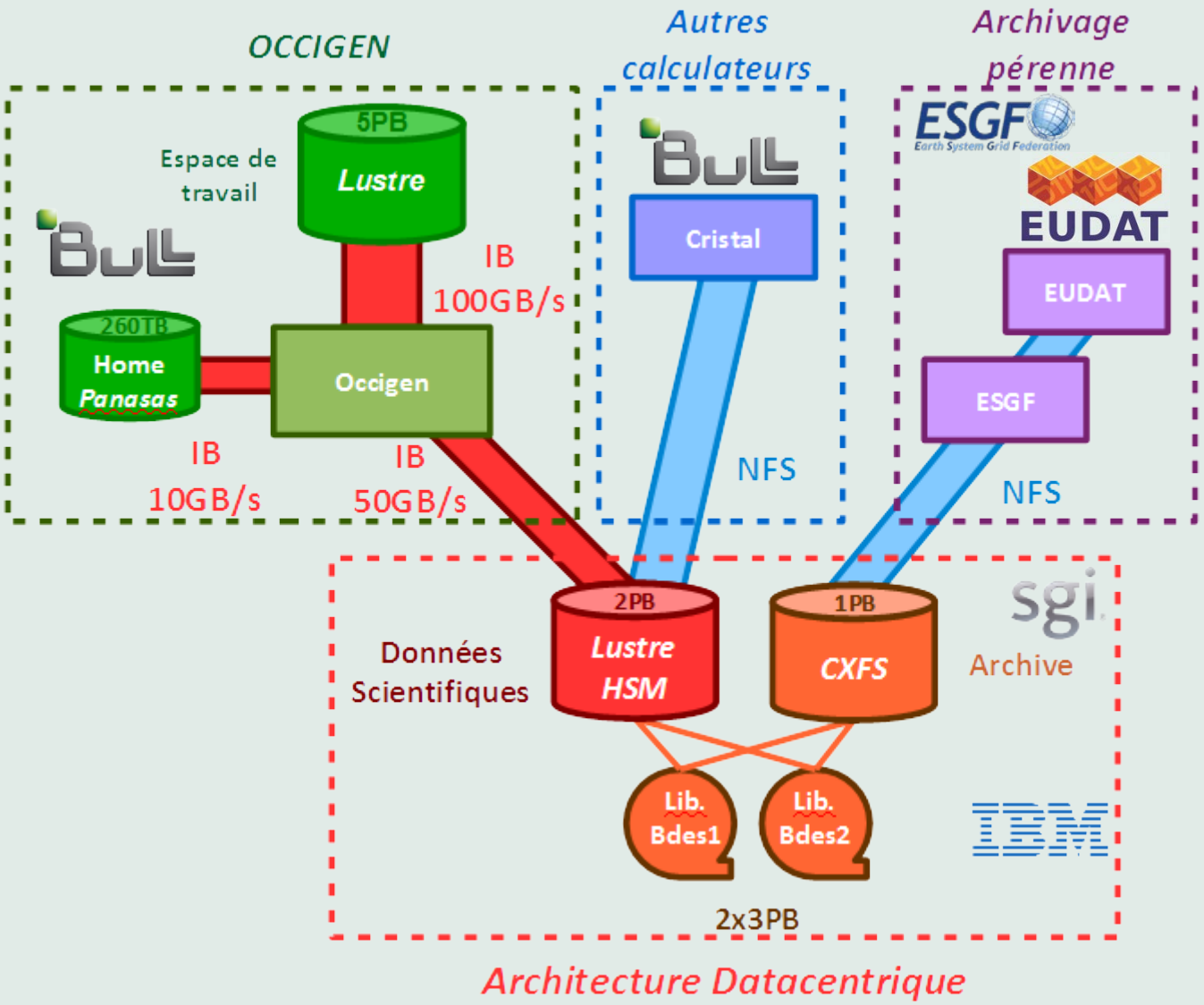
- **Lustre = Linux + Cluster**
- **Distribué sous licence GPL v2**
- **Permet un passage à l'échelle (scalabilité) :**
  - Utilisable sur plusieurs dizaines de milliers de nœuds,
  - Capacité de plusieurs dizaines de péta-octets,
  - Avec un niveau de performance de débit jusqu'à 1TB/s
  - Sans altération de la sécurité de l'ensemble.
- **Déployé sur la plupart des calculateurs du TOP500**
  - La moitié du TOP10
  - Les 2/3 du TOP100

lustre®
File System

L'historique



L'environnement de stockage



La fiche technique – Lustre

- **Réseau**
  - InfiniBand FDR (56GB)

- **Disques**
  - 20 x NetApp LSI E5500 (data)
    - 1200 disques: 3To@7200rpm , SAS – RAID 6
  - 1 x NetApp LSI E5500 (metadata)
    - 30 disques: 500 Go@10000rpm, SAS – RAID 6

- **Serveurs**
  - 2 x MDS, 1 cellule HA de 2 serveurs (actif/passif)
  - 20 x OSS : 5 cellules HA de 4 serveurs (actif/actif) – 240 OSTs

- **Logiciels**
  - BULL-PFS : *Community Lustre :* 2.5.3.90 (RHEL 6), Shine/HAShine
  - Migration en cours → *Intel Lustre :* IEEL 3.0.1 (RHEL 7) + Shine/HAShine

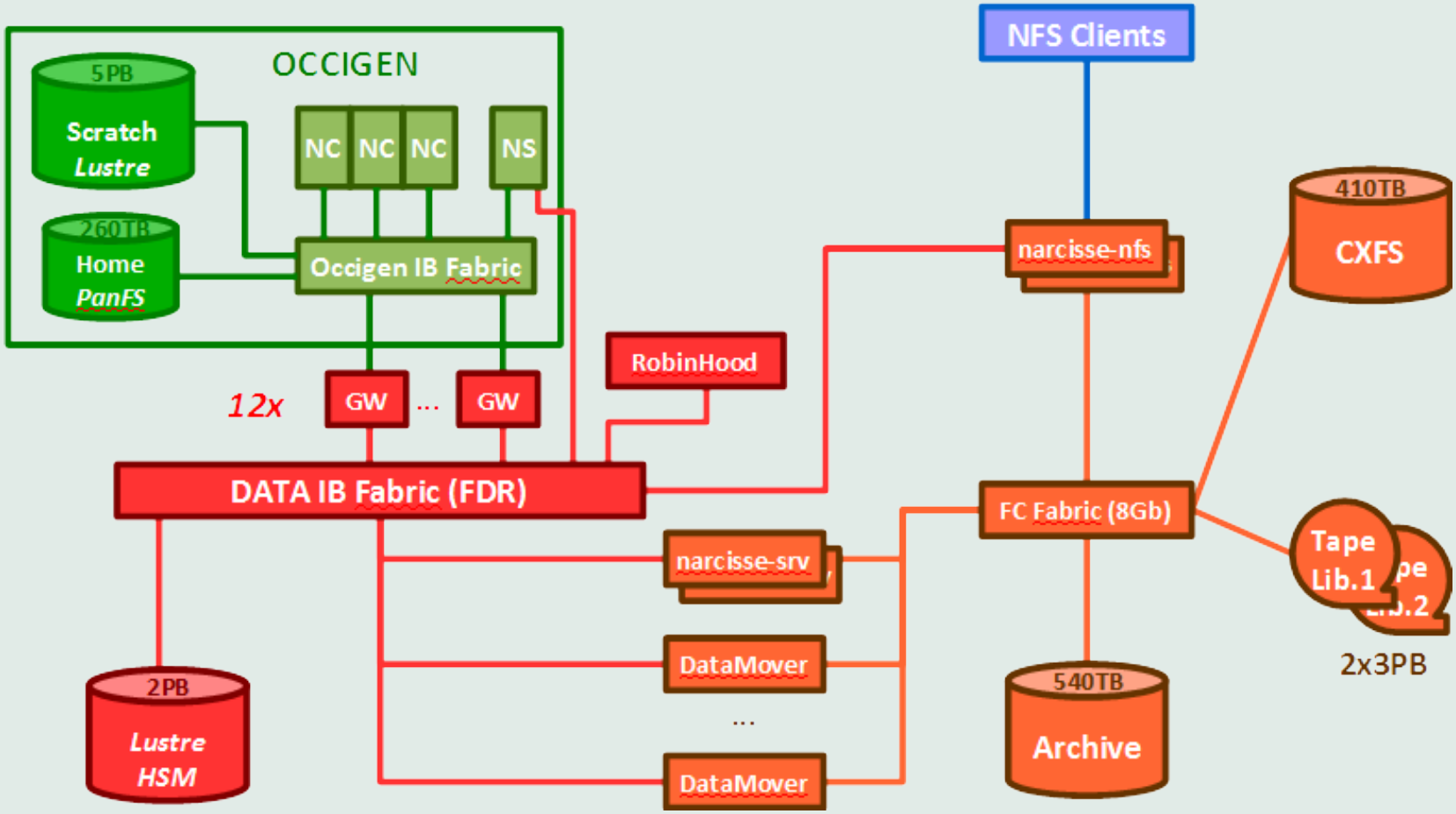
La fiche technique – Lustre HSM

- **Réseau**
  - InfiniBand FDR (56GB)
  - 12 Passerelles LNET (Lustre Network)
- **Disques**
  - 2 x DDN SFA 12k
    - 420 x disques 3TB SAS
- **Serveurs**
  - 2 x MDS
  - 12 x OSS – 84 OST
- **Logiciels**
  - Lustre 2.5.42.8 (IEEL v2) – RHEL 6

La fiche technique pDMF & CXFS

- **Réseau**
  - 2 x switch FC (Brocade 300e - 24 x 8GB)
- **Disques**
  - 2 x IS5500 (NetApp E5400)
    - 120 x disques 2TB SAS
- **Serveurs**
  - 2 x serveurs CXFS & DMS (HA)
  - 3 x datamovers DMF
  - 2 x serveurs NFS

Le diagramme fonctionnel



La volumétrie - Lustre

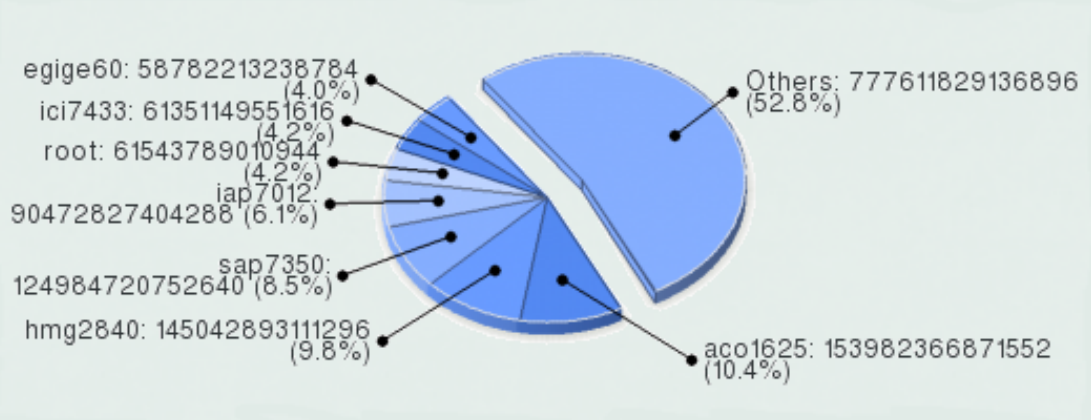
Groupe	Espace utilisé	Nombre fichiers
aco1625	140.05 TB	690 459
hmg2840	131.92 TB	5 156 019
sap7350	113.67 TB	2 796 141
iap7012	82.28 TB	8 520 013
ici7433	55.8 TB	246 487
egige60	53.46 TB	51 366
autres	819.45 TB	100 698 445

Robinhood

Policy Engine

Total : 1.5 PB

118 millions de fichiers



La volumétrie – Lustre-HSM

Etat	Espace utilisé	Nombre fichiers
synchro	1.54 PB	9 751 515
released	275.77 MB	564 772
unknown	2.17 GB	452 759
new	321.07 GB	152 484
modified	21.51 GB	536

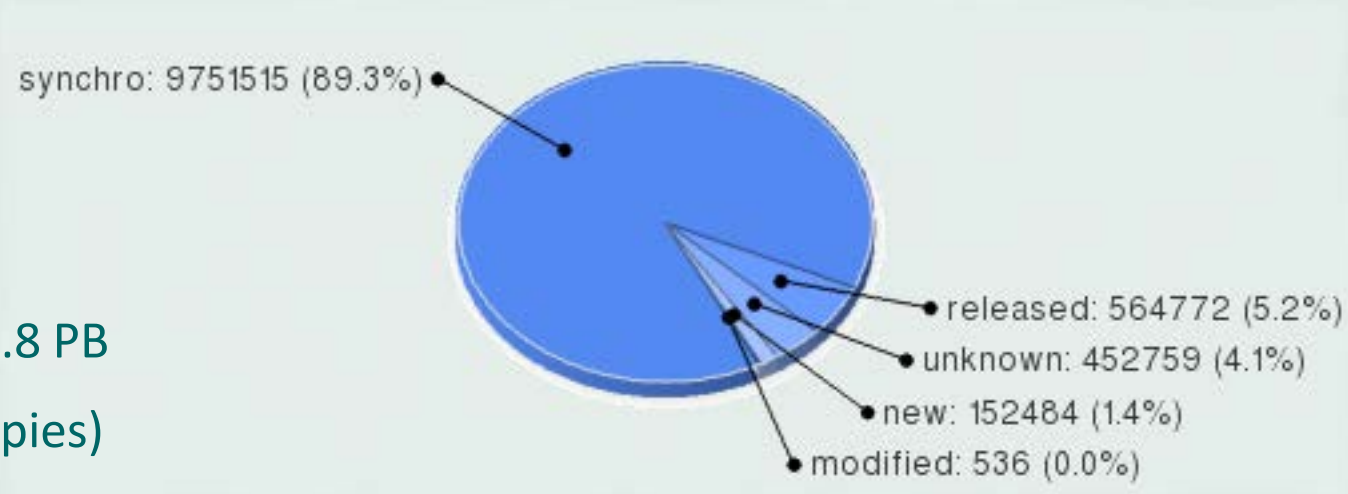
Robinhood

Policy Engine

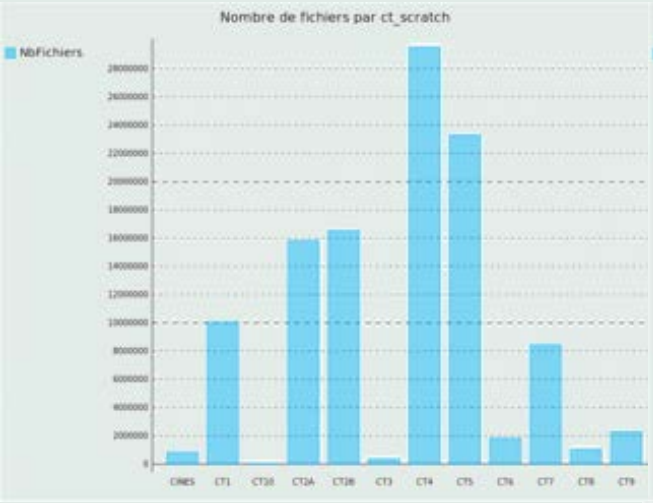
Total : 2.8 PB

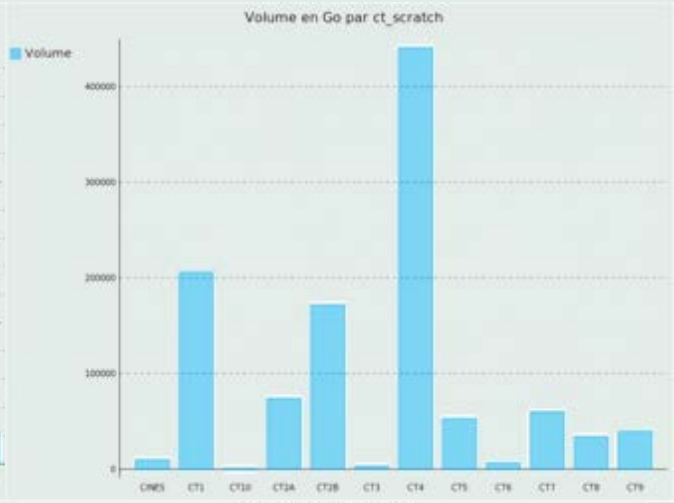
Dont Lustre-HSM 1.8 PB

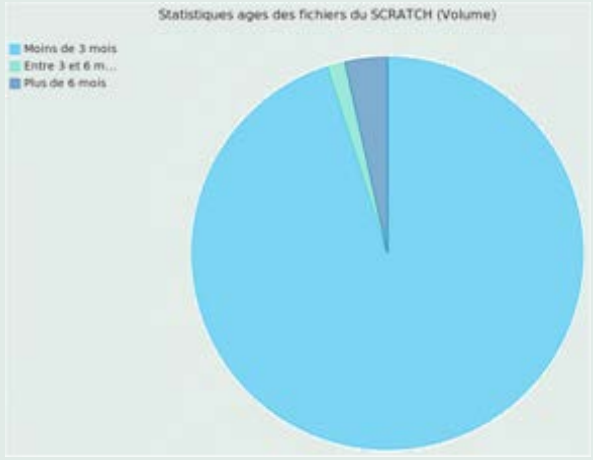
Bandes : 1 PB (2 copies)

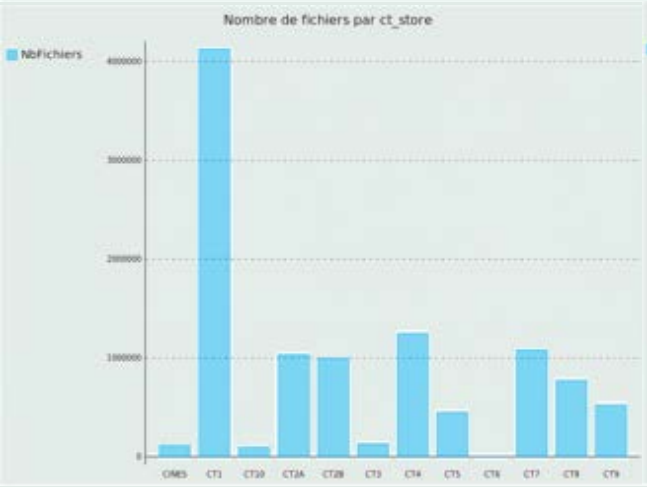


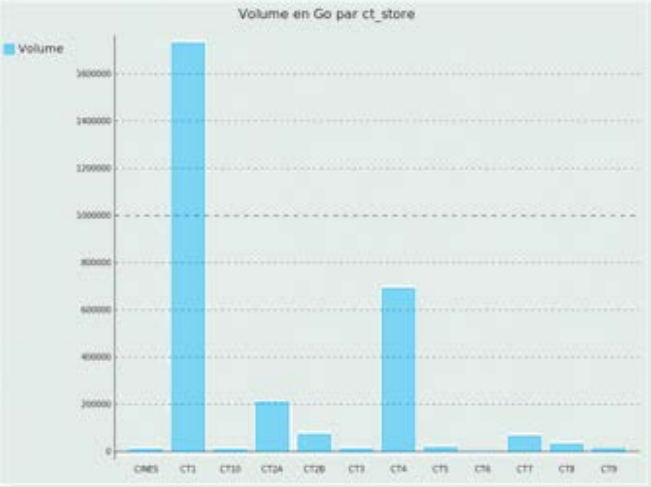
Les statistiques des espaces de stockage

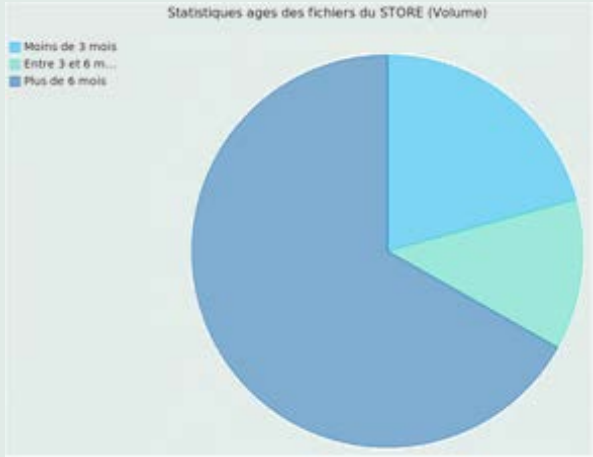












/scratch	1.5 PB	118 millions de fichiers
/store	2.8 PB (1 PB offline)	11 millions de fichiers

- **Mécanisme pour la gestion automatique des espaces de stockage**
  - Développé en interne
  - Implémenté pour tous les espaces utilisateurs : /home, /scratch, /store
  - Vérification de la fréquence d'utilisation (nombre de fichiers, volume, période, etc.)
  - En cas de dépassement (via SLURM) :
    - Les travaux en cours peuvent s'achever ;
    - Les travaux en attente sont bloqués dans une file avec un statut spécial
    - Les nouvelles soumissions de travaux sont rejetées
  - Les utilisateurs peuvent débloquer ces contraintes par une gestion rigoureuse de leurs espaces ;
  - Quotas par défaut : exceptions à la demande (via ticket par l'utilisateur)
- **Future interface avec RobinHood**
  - Vérification de la qualité : âge fichiers, etc.
  - Plus de statistiques disponibles ;
  - Suppression de fichiers à implémenter

Les points forts / faibles

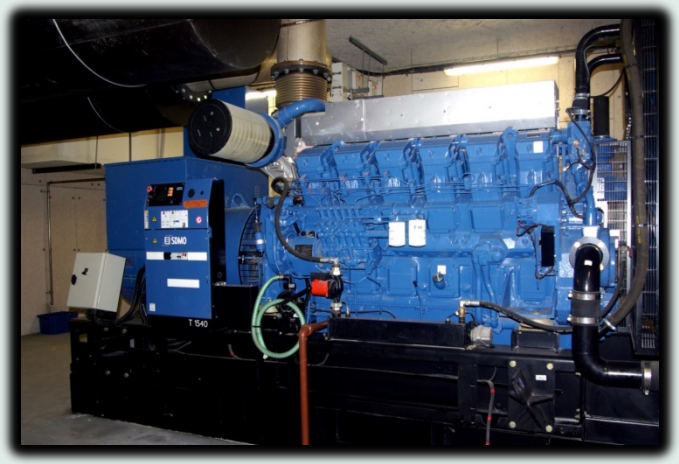
	Avantages	Inconvénients
Lustre	Performances et rapport coût/performance	Complexité d'analyse en cas de dysfonctionnement
	Grande scalabilité (client / serveur)	Tuning délicat du HA
	Environnement d'administration centralisée Shine en ligne de commande	Pas de Haute disponibilité MDS actif/actif
	Haute disponibilité (OSS + HAShine)	
	Fonctionnement stable (depuis version 2.x)	
	Tuning performances fichiers par l'utilisateur	
Lustre-HSM	Cohérence Lustre de bout en bout (stockage rapide, stockage sécurisé)	Absence de fonctionnalité de restauration « standard »
	Expertise interne développée	Retard 6 mois dans la migration Lustre HSM (bugs copytool)
	Fonctionnement HSM stable	IHM Intel pour l'administration HSM non supportée par l'intégrateur

Merci de votre attention

Plus d'information: http://www.cines.fr/

olivier.rouchon@cines.fr

Groupes froids



Groupe électrogène

Adduction réseau

Nouvelle salle machine : 600 m²