energie atomique • energies alternatives

Lustre au CEA
Retour d'expérience

Aurélien DEGREMONT
Thomas LEIBOVICI
CEA/DAM/DIF

Lustre : Généralités

energie atomique • energies alternatives

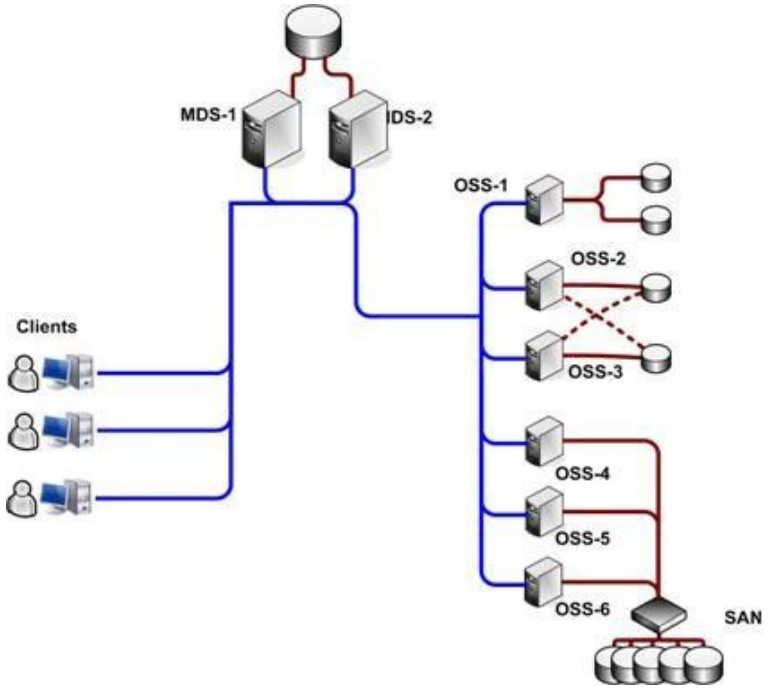
- Lustre : système de fichiers parallèle
  - Conçu pour des performances et une extensibilité maximale
- Développé par CFS -> Sun -> Oracle
- Forte présence dans le TOP 500 (autant que le TOP 10)

- OpenSource (GPL)
- Fonctionne sous Linux.
  - Plateformes officiellement supportées : RHEL, OEL, SUSE (client)
  - Fonctionne aussi sur Debian/Ubuntu.

Lustre : Architecture

energie atomique • energies alternatives

- **Gestion distribuées des données**
  - 1 serveur de gestion de meta-données: MDS
  - N serveurs de stockage des données des fichiers: OSS
  - Données exposables via plusieurs serveurs.



- **Stockage des données**
  - Données réparties en stripe (RAID0) sur un ou plusieurs OST
    - Stripping configurable.
  - Backend ext4+patch
  - Quotas

Retour d'expérience (1/3)

energie atomique • energies alternatives

- Performances
  - Bonne scalabilité, débit, Iops
  - Un seul serveur de métadonnées
    - Suffisant car accès optimisé (par design)

- Technologies réseau
  - Support natif de Ethernet, Infiniband, Myrinet, ...
  - Pas d'aggregation de liens par Lustre

- Installation / configuration
  - ✓ Installation standard et facile (RPMs)
  - ▪ Nécessite un kernel patché sur les serveurs
    - ✓ RPM du kernel patché également fourni
  - ▪ La configuration / le tuning peuvent être plus ardus
  - ✓ Des outils libres existent pour faciliter la tâche (*Shine*)

Retour d'expérience (2/3)

energie atomique • energies alternatives

- **Failover / tolérance aux pannes**
  - Résistant à une panne serveur: mécanismes de *recovery*.
  - Heartbeat, RedHat Cluster Suite, ...
  - Beaucoup peuvent être évitées par l'utilisation de *dm-multipath*

- **Administration**
  - Debug : compréhension des logs, dmesg sur les serveurs et/ou les clients.
    - Nécessite une expertise.
  - fsck long
    - ✓ Peut être parallélisé avec *Shine*
    - ➢ Ne sera plus nécessaire avec les OST en ZFS (actuel: ext4)

Retour d'expérience (3/3)

energie atomique • energies alternatives

- **Stabilité**
  - Amélioration constante de version en version
  - Nouvelles priorités de développement :
    - Lustre 1.x : #1 perf #2 features #3 stabilité
    - Lustre 2 : #1 stabilité #2 perf #3 features

- **Export NFS/CIFS**
  - Pas intégré
  - Via un noeud client avec *nfsd*, *samba*
  - ✓ Export par plusieurs serveurs NFSv4 en parallèle à partir de Lustre 2 (avec *nfs-ganesha*)

- **Support**
  - Gratuit, via bugzilla
  - OpenSource, communauté de grands sites HPC
  - Solutions clef en main: Oracle (Lustre+HW)

energie atomique • energies alternatives

Lustre au CEA

Le besoin : TERA 100

energie atomique • energies alternatives

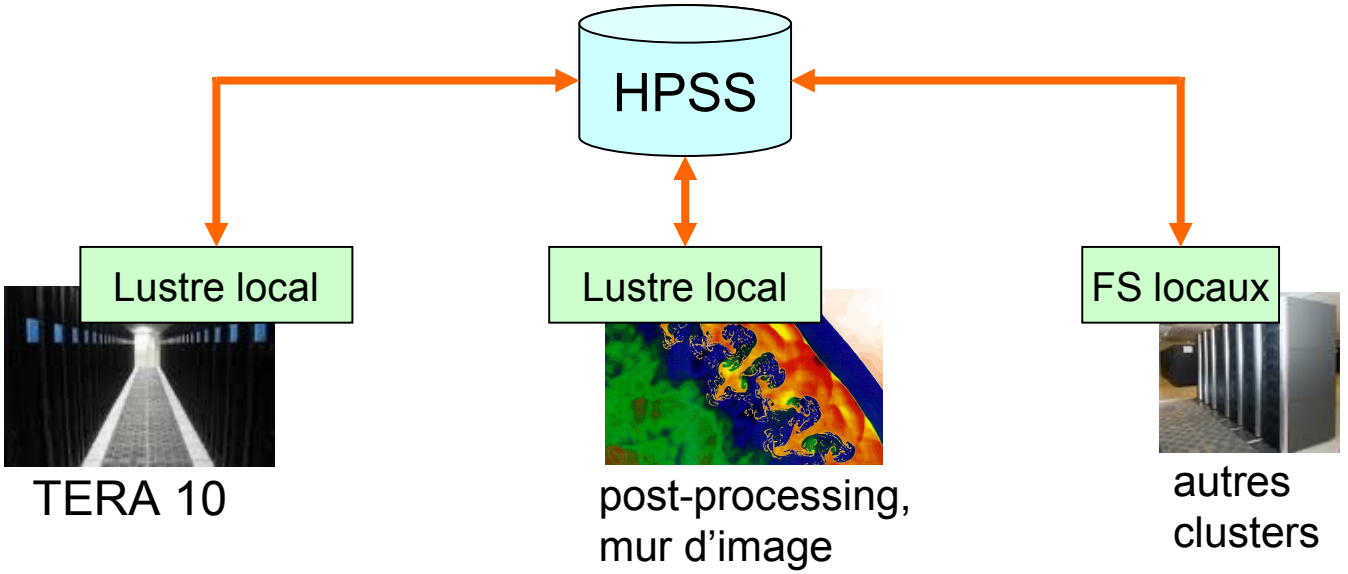
- **4300 Noeuds**
- **140 000 coeurs Nehalem-EX**
- **1,25 Petaflops**

- **Mémoire :**
  - 300 To
- **Stockage**
  - 20 Po
  - 500 Go/s
- **Interconnexion Infiniband QDR**

Architecture de Stockage TERA 10

energie atomique • energies alternatives

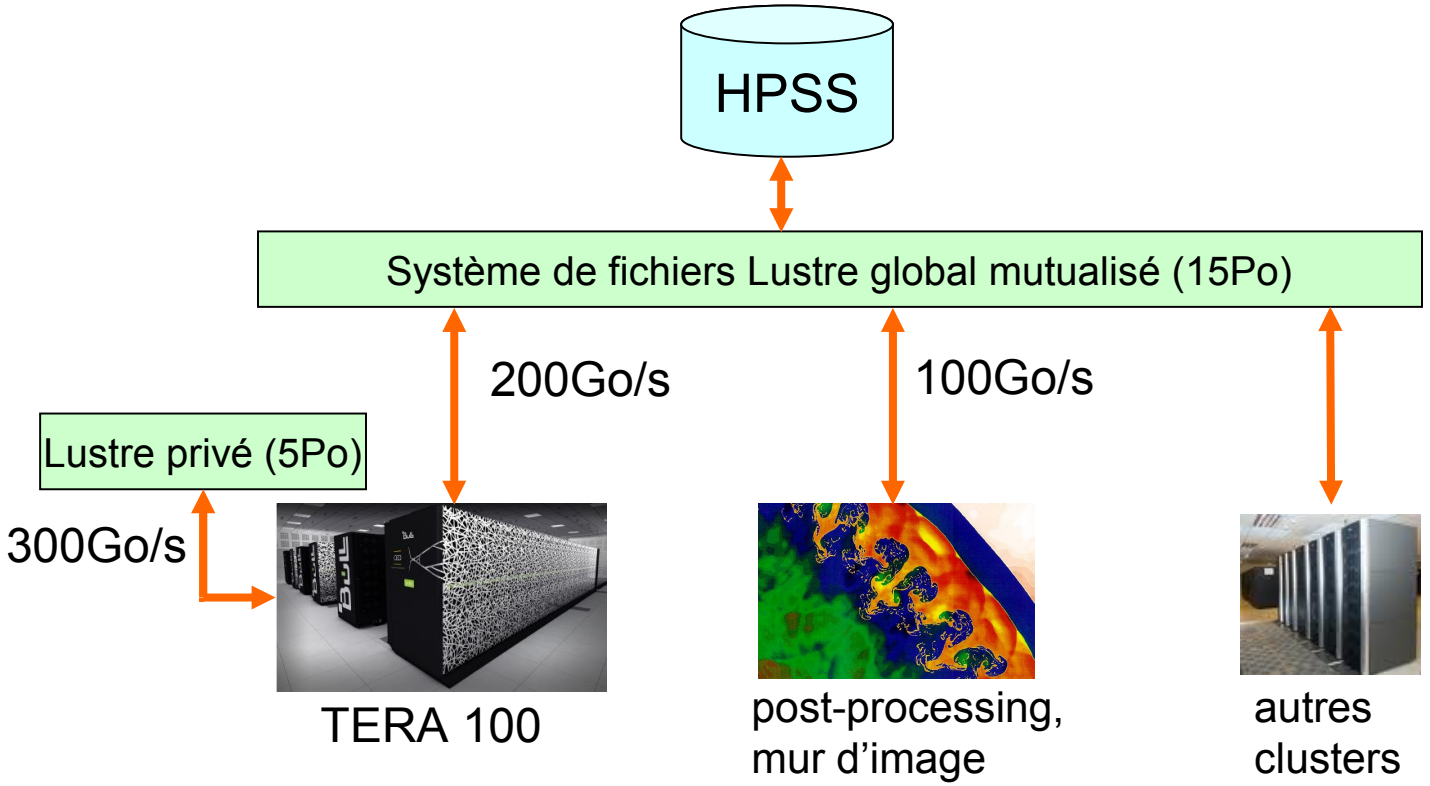
- Chaque cluster a son propre système de fichiers parallèle privé
- Transit des données par HPSS pour passer d'un cluster à l'autre



Points négatifs :

- Va et vient de données entre les clusters via HPSS
  - Latence d'accès aux données après calcul (post-processing)
  - Gaspillage de bande passante
- Nombreux systèmes de fichiers à administrer

Architecture de Stockage TERA 100



- Données temporaires, protections/reprises :
  - Système de fichiers Lustre privé
  - 300 Go/s
- Données de résultats:
  - Système de fichiers global de 15Po
  - 200Go/s vers TERA100 (100Go/s visu et autres clusters)
  - Zero-copie pour la visualisation

energie atomique • energies alternatives

Les outils spécifiques

- Administration Lustre à grande échelle
- Gestion des données
- Hétérogénéité
- Sauvegarde, stockage à long-terme

Administration à grande échelle : Shine

energie atomique • energies alternatives

- **Besoins d'administration spécifiques à TERA100 :**
  - +4300 clients Lustre
  - 64 OSS (Lustre privé) + 66 OSS (Lustre global)

- **Outil : Shine**
  - Configuration et l'administration de Lustre à grande échelle
  - Collaboration CEA/Bull
  - Open Source

- **Principales fonctionnalités :**
  - Administration et configuration centralisée
  - Actions en parallèle (mount, fsck…)
  - Simplifie la configuration de Lustre

- **http://lustre-shine.sf.net**

Gestion des données : Robinhood

energie atomique • energies alternatives

- **Besoins :**
  - Surveiller/auditer le contenu d'un système de fichiers
    - Volume par utilisateur, par répertoire, …
    - Plus gros fichiers, plus vieux fichiers, …
    - Fichiers par OST, par utilisateur, …
  - Appliquer des politiques
    - Purge des fichiers inutilisés, blacklist/whitelist
  - Sur des centaines de millions de fichiers
    - 'find', 'du' quasi inutilisables…

- **Outil : Robinhood**
  - Collecte des informations du FS dans une base de données
    - Interrogeable à volonté sans parcours du FS
  - Définition de politiques de purge/migration
    - Basées sur les attributs du fichier (chemin, taille, propriétaire, age…)
    - Whitelists

Gestion des données : Robinhood (suite)

energie atomique • energies alternatives

- **Robinhood (suite)**
  - Collecte d'informations du FS
    - Lustre 2: pas de scan nécessaire (ChangeLogs)
    - Autres FS: scan parallèle, mais vitesse non critique (DB persistante)
  - Développé par le CEA
  - OpenSource (CeCILL)

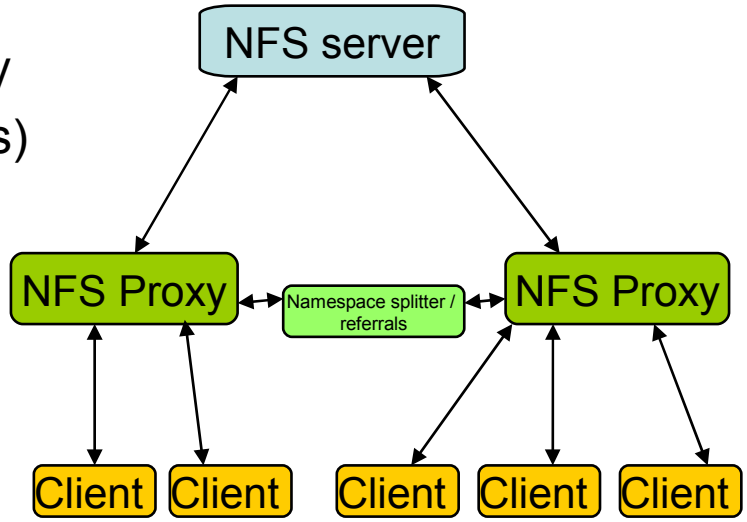
- **http://robinhood.sf.net**

Export NFS via NFS-GANESHA

- Besoins liés à l'hétérogénéité :
  - Accéder à Lustre sur des architectures non supportées
    - Ex : Montage sur un parc de stations hétérogène…
  - Via un protocol standard : NFS

- Serveur NFS-GANESHA
  - Serveur NFS en User-Space, massivement multi-thread
  - Capable d'exporter un filesystem Lustre 2.x (entre autres)
  - Utilisé en tant que serveur NFS proxy, de répartir la charge entre plusieurs proxy
  - Export en NFSv3, NFSv4 (pNFS en cours)
  - Développé par le CEA
  - OpenSource

- http://nfs-ganesha.sf.net

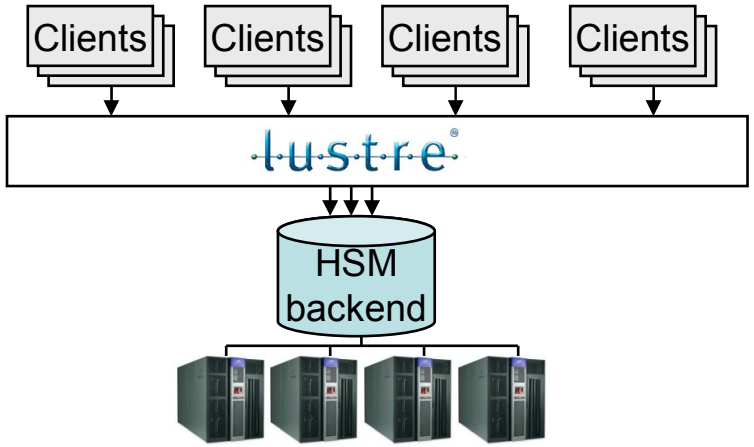


Stockage à Long-Terme : Lustre-HSM

energie atomique • energies alternatives

- Besoins :
  - Migration/rapatriement automatique des données de Lustre avec un backend (HSM, …)
  - Transparent pour l'utilisateur (tous les fichiers restent visibles dans Lustre)

- Solution : Lustre-HSM binding
  - Fonctionnalité qui sera intégrée à Lustre (2.x)
  - Support de nombreux backend
    - HPSS, POSIX, Enstore
    - A venir : DMF, TSM, …
  - Collaboration CEA/Oracle
  - Statut : fin d'implémentation



energie atomique • energies alternatives

Bilan

Bilan

- **Lustre est un système de fichiers très performant**
- **Ayant cependant une complexité relative**
- **Mais atténuée grâce :**
  - à une communauté importante
  - au support d'Oracle
  - des outils open-source comme Shine, Robinhood ou Ganesha
- **C'est le système de fichiers de référence des plus puissantes machines au monde, mais utilisé sur des configurations de plus en plus petites.**

Questions ?