Systèmes de fichiers distribués : comparaison de GlusterFS, MooseFS et Ceph avec déploiement sur la grille de calcul Grid'5000.

JF. Garcia, F. Lévigne,
M. Douheret, V. Claudel

30 mars 2011

Table des Matières

Présentation du sujet

Comparaison de systèmes de fichiers distribués :

- Système de fichiers (FS) : façon de stocker, organiser des informations dans des fichiers sur une mémoire secondaire (CD-ROM, disque dur,...)
- Système de fichiers distribué :
  - éclaté sur plusieurs serveurs
  - disponible depuis plusieurs clients

Le Grid'5000

- Infrastructure distribuée dédiée à la recherche
- 11 sites, dont 9 en France

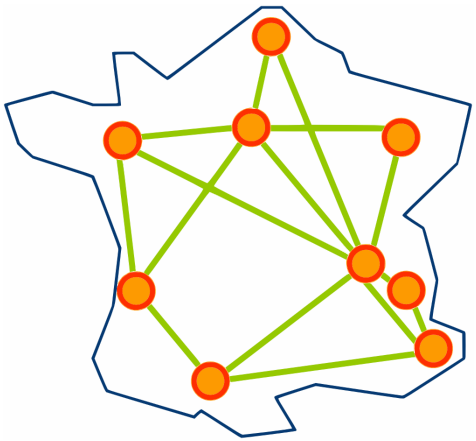


Figure: Les sites français du Grid'5000

Travailler sur le Grid'5000

- Connexion au « frontend » par SSH
- Réservation de nœuds, pour un certain temps
- Déploiement d'image (OS)

Astuce :

Possibilité d'effectuer une réservation à l'avance, suivit par l'exécution d'un script

Présentation de NFS

- Network File System
- Développé par Sun Microsystem en 1984
- Partager des données par le réseau
- Méthode standard de partage entre machines Unix

Aspect technique

- NFS et le protocole non connecté UDP
- Depuis la version 3, possibilité d'utiliser TCP
- Versions NFS définies dans différentes RFC
- Ensemble du protocole repensé pour NFSv4 :
  - meilleur gestion de la sécurité
  - meilleur gestion de la montée en charge
  - système de maintenance simplifié
  - support des protocoles TCP (par défaut) et RDMA

Mise en place

- Installation des paquets nfs-common et nfs-kernel-server
- Implémentation d'un fichier exports dans /etc
- Montage du partage sur les clients à l'aide de « mount »

Pour NFSv4 :

Des options supplémentaires sont à définir dans /etc/exports et le type de protocole doit être spécifié lors du montage sur les clients.

Présentation de GlusterFS

- Licence GPLv3
- Se base sur FUSE (Filesystem in UserSpacE)
- Capacité pouvant atteindre plusieurs petabytes (1000 To)
- Structure simple, deux éléments logiciels : serveur et client
- Supporte plusieurs protocoles de communications (TCP/IP, InfiniBand)

Mise en place

- Un serveur maitre : paquet glusterfs-server
- x serveurs « normaux »
- x clients : glusterfs-client

Note :

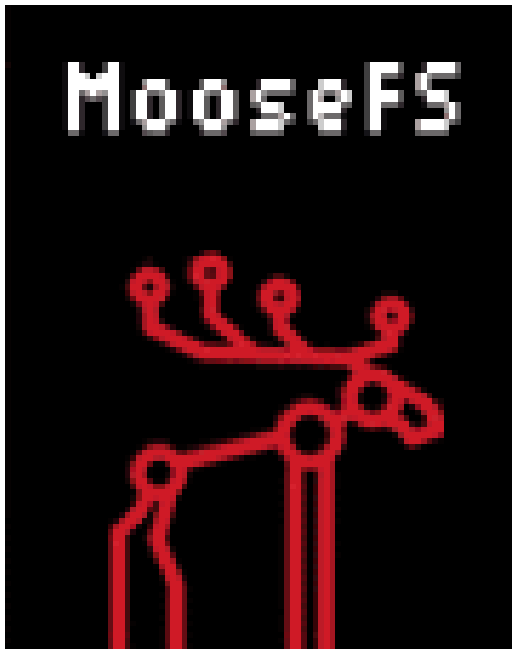
Les serveurs doivent avoir un répertoire dédié au partage

Mise en place (2)

- A partir du serveur maitre :
  - génération des fichiers de configurations (commande prévue)
  - envoie de fichiers aux serveurs, et aux clients
- Démarrage des serveurs
- Montage du volume par les clients

Difficultés rencontrées

- Droit d'écriture des clients
- Utilisation d'InfiniBand


MooseFS

Présentation de MooseFS

MooseFS (Moose File System) est un système de fichiers répartis à tolérance de panne, développé par Gemius SA.

- Licence GPLv3.
- Disponible pour Linux, FreeBSD, OpenSolaris et MacOS X.
- Respect de la norme Posix et l'utilisation de Fuse en espace client.
- Sa simplicité d'administration, de mise en œuvre et d'utilisation.
- Poubelle par défaut.
- scalable

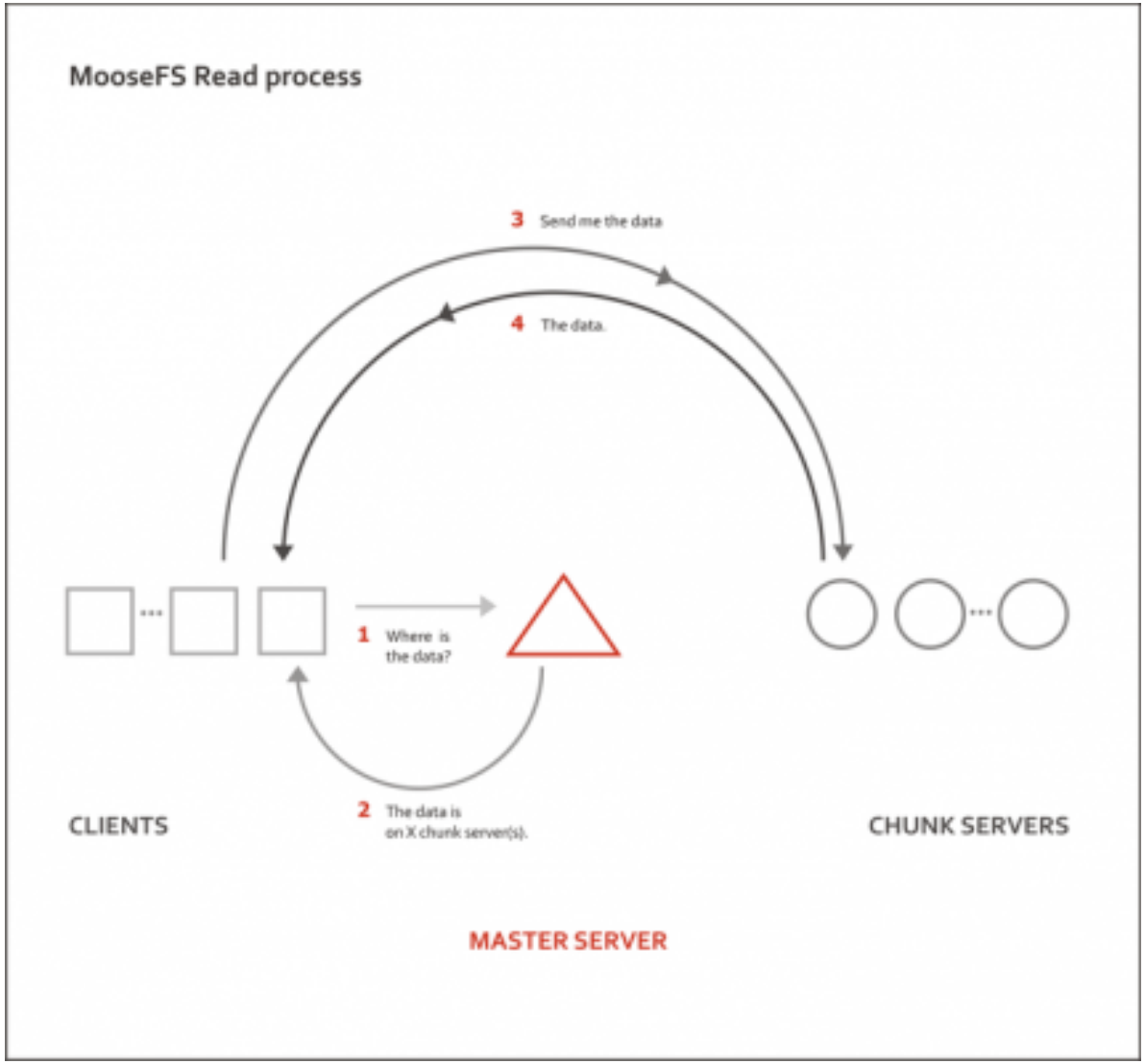
Architecture

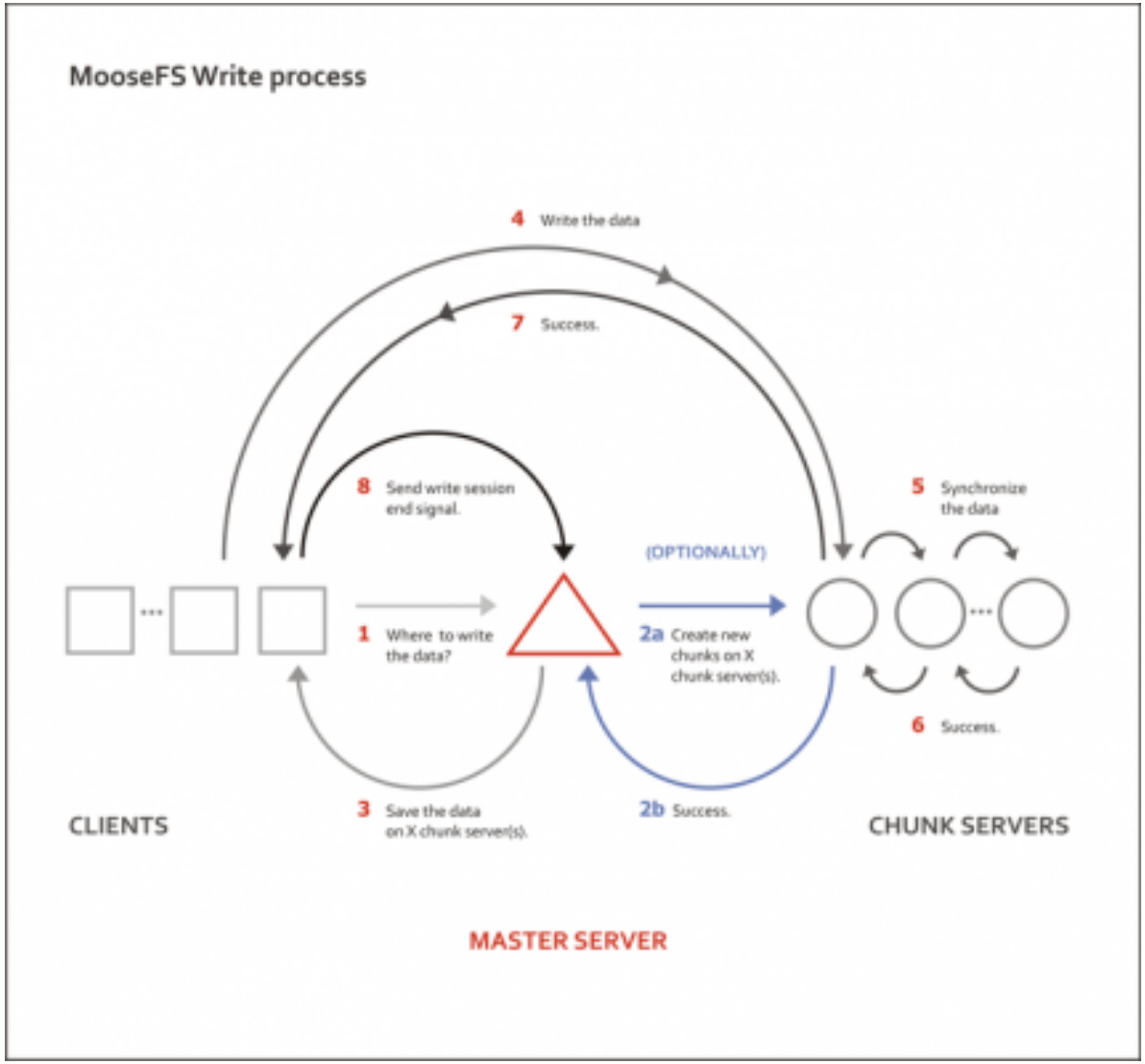
MooseFS est constitué de trois types de serveurs :

- Le Master Serveur
- Le Metalogger Serveur
- Le Chunck Serveur

Fonctionnalités

- Tolérance aux pannes
- Le système est réparti
- Répartition de charge
- Sécurité


MooseFS Read process
3 Send me the data
4 The data
1 Where is the data?
2 The data is on X chunk server(s).
CLIENTS
CHUNK SERVERS
MASTER SERVER


MooseFS Write process
4 Write the data
7 Success.
8 Send write session end signal.
5 Synchronize the data
(OPTIONALLY)
1 Where to write the data?
2a Create new chunks on X chunk server(s).
6 Success.
3 Save the data on X chunk server(s).
2b Success.
CLIENTS
CHUNK SERVERS
MASTER SERVER

Présentation de Ceph

- Licence LGPL
- Créé par Sage Weill en 2007
- Destiné aux très grands clusters
- But principal :
  - compatible POSIX
  - complètement distribué sans point de défaillance

Caractéristiques

- Robustesse
- Évolutivité transparente
- Déconseillé en production

Fonctionnement

Trois types distincts de démons :

- Moniteur de cluster
- Serveurs de métadonnées
- Serveurs de données

Moniteur

- Configuration
- État du cluster
- Gestion des clients

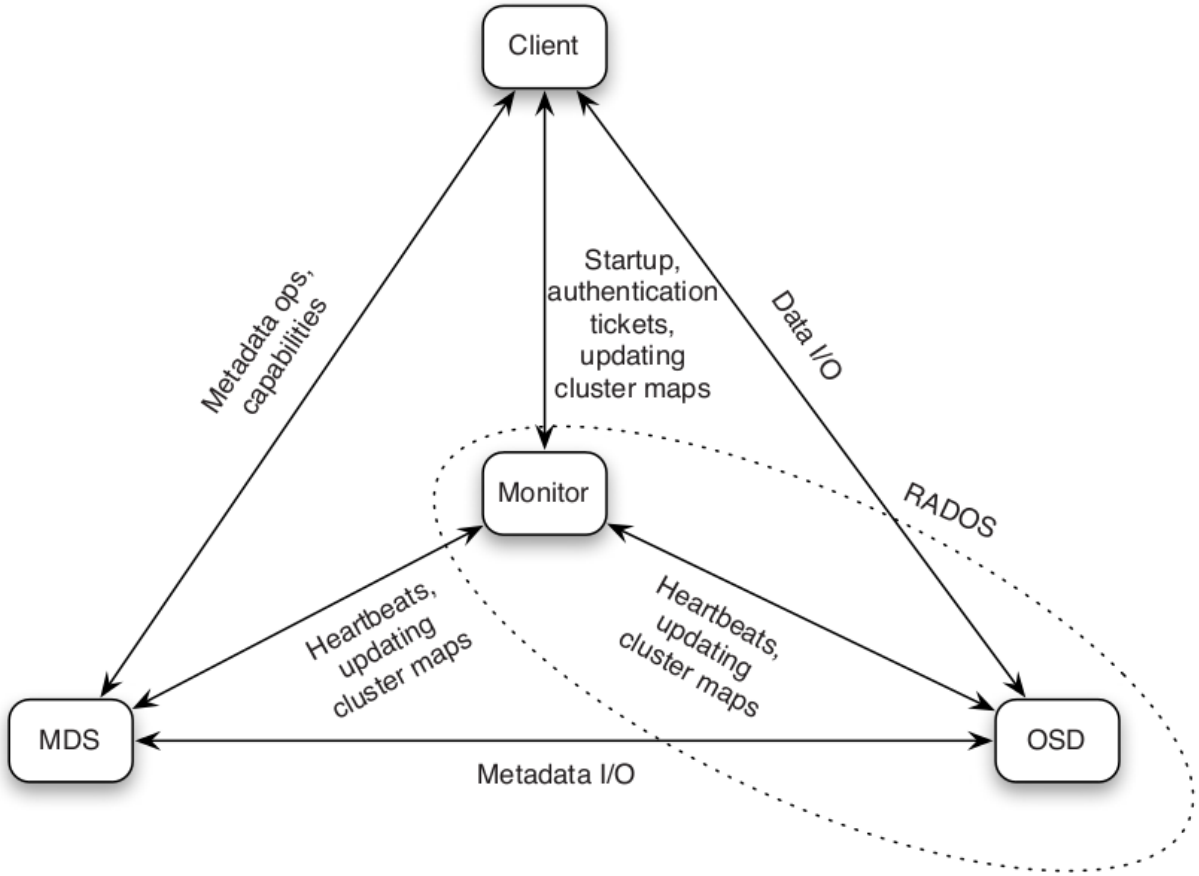
Serveurs de métadonnées

- Cache cohérent et distribué
- Plusieurs serveurs = équilibrage de charge

Serveurs de données

- Découpage des données
- Réplication = tolérance aux pannes

Echanges de données



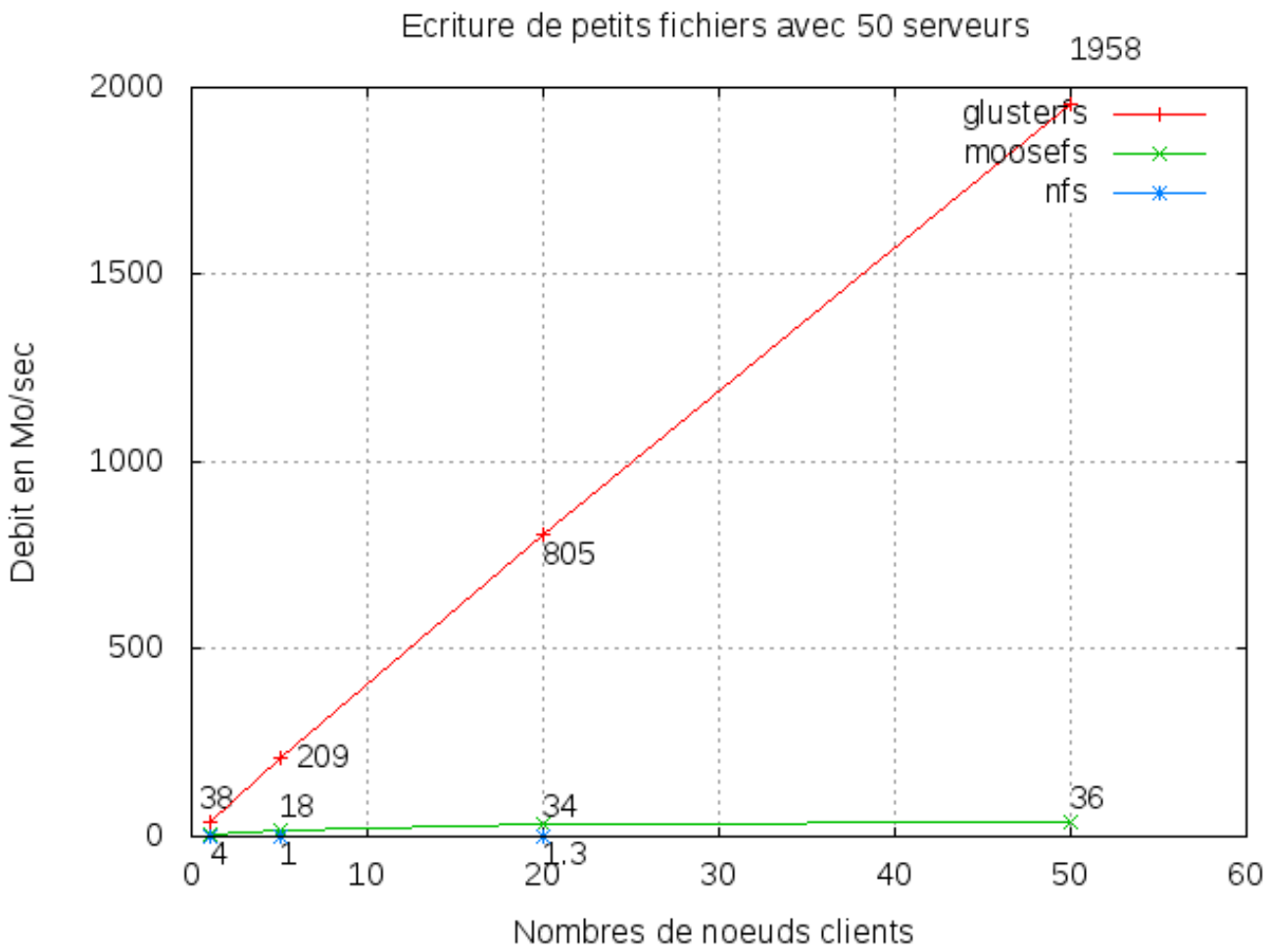
Difficultés rencontrées

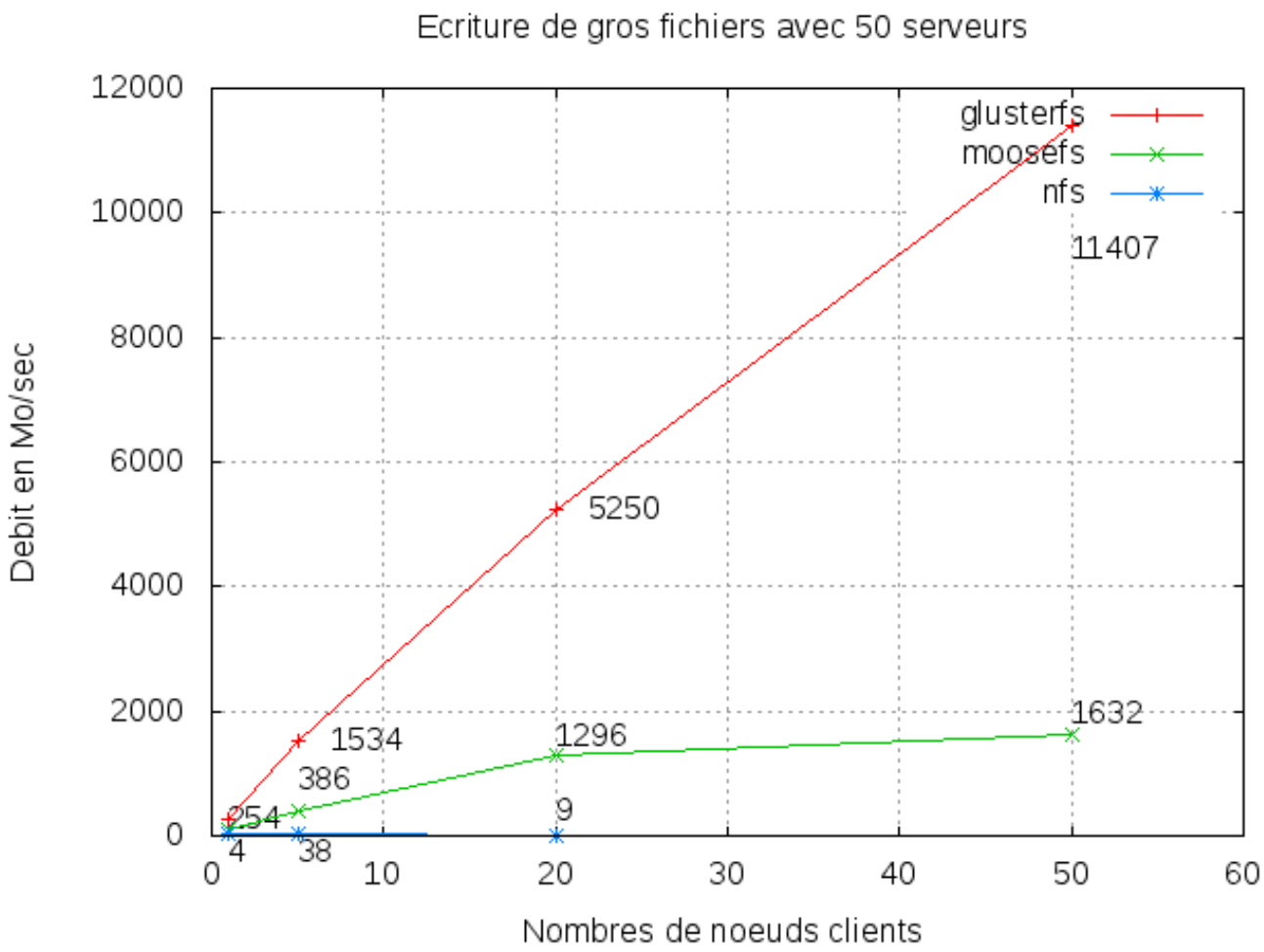
- Documentation minimaliste
- Fichier authentification

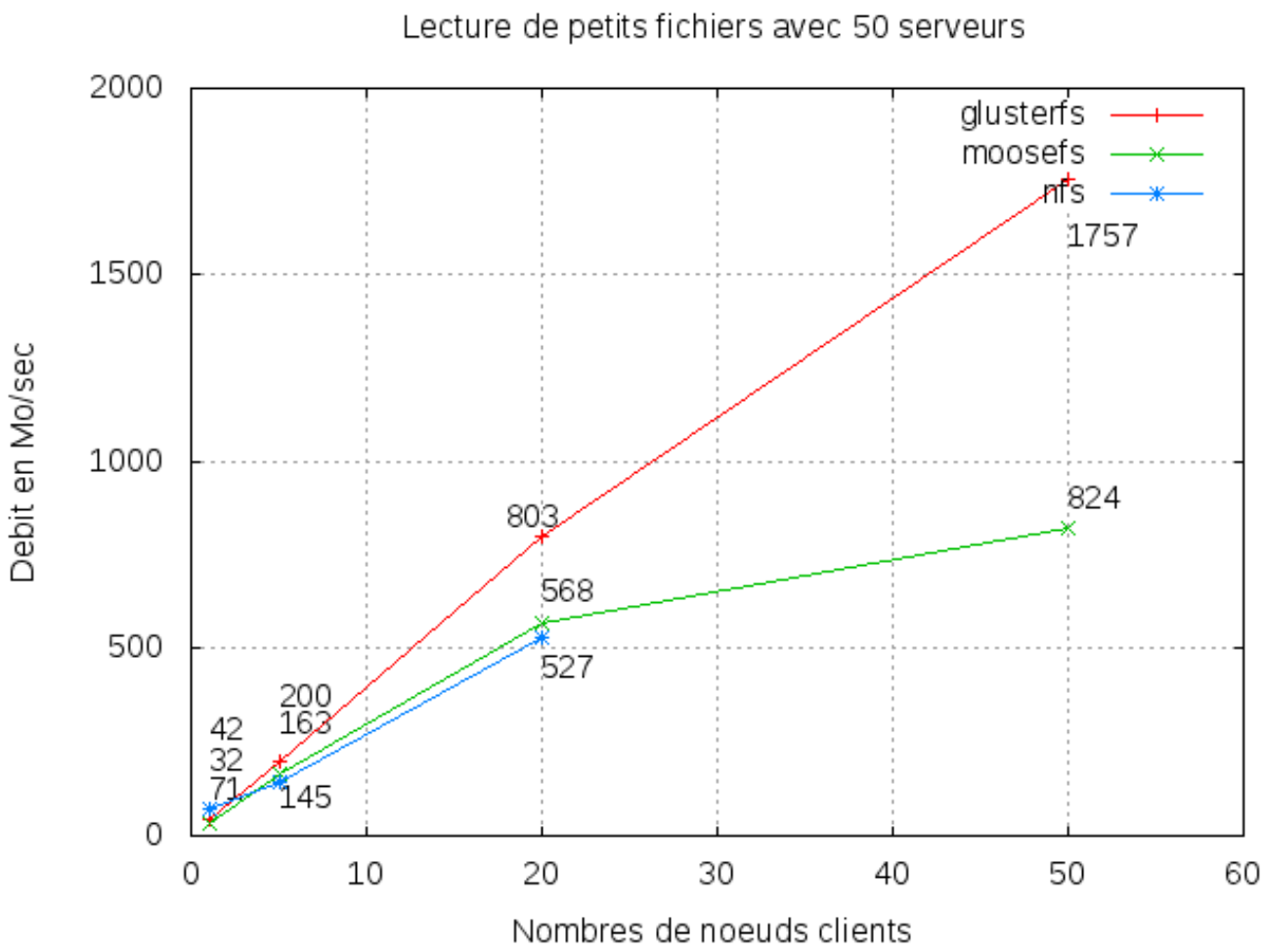
Benchmark

Actions simultanées sur plusieurs clients :

- Écriture de petits fichiers
- Écriture de gros fichiers
- Lecture de petits fichiers
- Lecture de gros fichiers


Ecriture de petits fichiers avec 50 serveurs
Debit en Mo/sec
Nombres de noeuds clients
glusterfs
moosefs
nfs
2000
1500
1000
500
0
0
10
20
30
40
50
60
1958
805
209
38
18
4
1
34
1.3
36


Ecriture de gros fichiers avec 50 serveurs
Debit en Mo/sec
Nombres de noeuds clients
glusterfs
moosefs
nfs
12000
10000
8000
6000
4000
2000
0
0
10
20
30
40
50
60
11407
5250
1534
1632
1296
386
254
9
4
38


Lecture de petits fichiers avec 50 serveurs
Debit en Mo/sec
Nombres de noeuds clients
glusterfs
moosefs
nfs
2000
1500
1000
500
0
0
10
20
30
40
50
60
1757
824
803
568
527
200
163
145
42
32
71

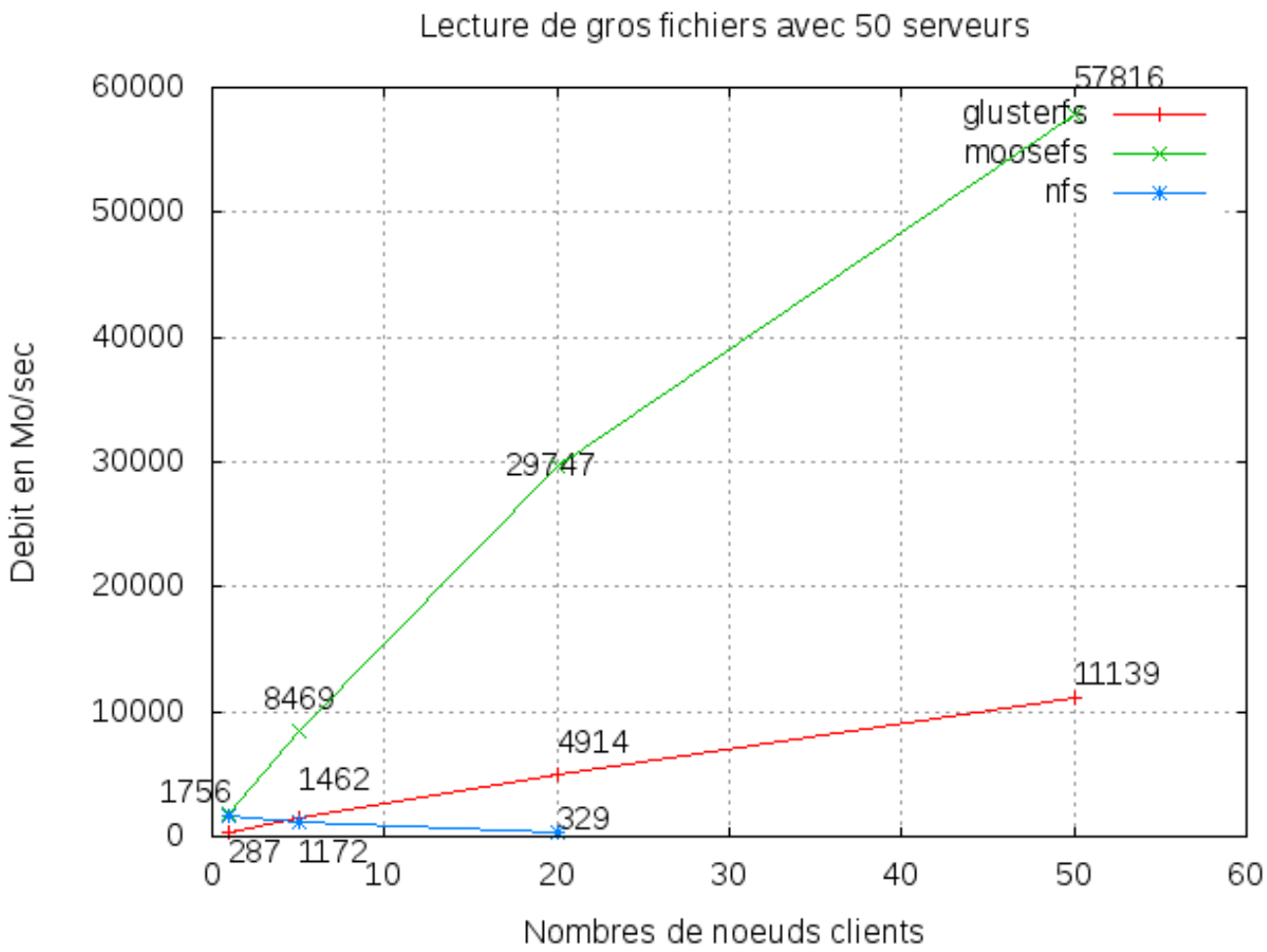

Lecture de gros fichiers avec 50 serveurs
Debit en Mo/sec
Nombres de noeuds clients
glusterfs
moosefs
nfs
60000
50000
40000
30000
20000
10000
0
0
10
20
30
40
50
60
57816
29747
11139
8469
4914
1462
1756
329
287
1172

Tableau comparatif

	Gluster	Moose	Ceph	NFS
Facilité de mise en place	++	+	+	++
Fiabilité	++	++	-	++
Sécurité, disponibilité des données	+	++	++	- -
Évolutivité	+	++	++	- -
Économe en taille disque	++	-	-	++

Difficultés rencontrées

- Prise en main du Grid'5000
- Partage du cluster
- Erreurs ponctuelles lors de déploiements
- Scripts de déploiements, benchmark : automatisation totale

Travail accompli

- Mise en place de systèmes de fichiers distribués
- Création de scripts de déploiements, et de benchmark
- Comparaison de ces systèmes

Expérience enrichissante

- Travail sur un cluster
- Niveau de technique important
- Documentations en anglais