ÉCOLE **NORMALE**
SUPÉRIEURE
DE **LYON**

JRES 2011

Le projet DiStoNet

DIStributed **STO**rage **NET**work

«Dupliquez, dupliquez, il en restera toujours quelque part ! »

Emmanuel Quemener
Loïs Taulelle

ÉCOLE **NORMALE**
SUPÉRIEURE
DE **LYON**

Une origine, le projet ARMUS

- **Quand** ? Idée de fin 2004...
- **Où** ? ENS-Cachan (oui, une « autre » ENS)
- **Qui** ? Votre serviteur ☺ !
- **Quoi** ? Architecture Répartie Mutualisée entre Utilisateurs pour le Stockage
- **Combien** ? Après le comment...
- **Pourquoi** ? Rationalisation des ressources
- **Comment** ? Restée au stade d'idée !

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Stockage distribué = original ? Pas spécialement...

- *« Distributed file system »*
  - Sur Wikipedia : 60 entrées
    - Si « Tolérance aux pannes » : 38
    - Si « Open Source » : 11
    - Si « en activité » : 8
  - En production : Coda, MooseFS, Gfarm, GlusterFS, Lustre, XTreemFS
  - En développement : CephFS, Hadoop
- *« Disk over network »*
  - 6 trouvés : NBD, DRDB, FCIP, FCoE, AoE, iSCSI

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Un renouveau : enquêtes besoins de stockage de 2010

- Plusieurs éléments de contexte :
  - Estimation de coûts « classiques » très importants
  - Volonté de « valorisation » technologique de grilles
  - Mise à disposition d'équipements déclassés
  - Expérimentation sur systèmes de fichiers récents
  - Nécessité de passer à l'échelle
  - Interrogation de la direction de l'ENS-Lyon
- Création du projet DiStoNet !
- Et puis, fin 2011, des inondations en Thaïlande...

ÉCOLE **NORMALE**
SUPÉRIEURE
DE **LYON**

Nos unités : le CBP & le PSMN

Centre Blaise Pascal

Un centre d'essais

et de formation

Pôle Scientifique de Modélisation Numérique

Une usine de production

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Contexte : des besoins croissants
Mais des solutions onéreuses...

- Enquête sur les moyens de stockage
  - Projection au Péta-Octet à 4 ans à l'ENS-Lyon
  - Estimation de prix pharaonique (en SAN standard)
- Coût du Téra-octet SAN : 500€/To
- Coût du Téra-octet « poste de travail » : 45€/To
- Rapport stockage sur hôte :
  - SAN : 200€ / disque, sur 48 disques, rapport 6
  - Poste de travail : 100€ / disque, rapport 10
- Augmentation prix des disques durs (x2~3 en 2011-11)

ÉCOLE **NORMALE**
SUPÉRIEURE
DE **LYON**

Contexte : des ressources inexploitées

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Contexte : des ressources inexploitées

- Grappe de calcul scientifique
  - Disques systèmes de 2x146 Go
- Postes de salle libre service
  - Disque interne de 250 à 500 Go
- Station de travail
  - Disque interne de 1 To
- Poste de bureautique
  - Disque interne de 250 Go
- Ordinateur portable
  - Disque interne de 500 Go (ou 128 Go en SSD)

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Contexte : Taux d'utilisation & Ressources inexploitées

- Humain : 35 h/semaine : 1607 h/an
- Machine 10h/j moins WE : 2612 h/an ~ 61% de TU
- Machine 24h/24 : 8766 h/an ~ 18% de TU
- Dans l'usage courant :
  - Processeur rarement à 100%, mémoire parfois...
  - Services distants : navigateur ou clients lourds
  - Dossiers distants : « disque » sur le réseau
  - Profil itinérant : portabilité d'un poste à l'autre
  - **Finalement, un disque local à des fins « personnelles »**

ÉCOLE **NORMALE**
SUPÉRIEURE
DE **LYON**

Le stockage - les deux approches mode « bloc » & mode « fichier »

Le mode « bloc »

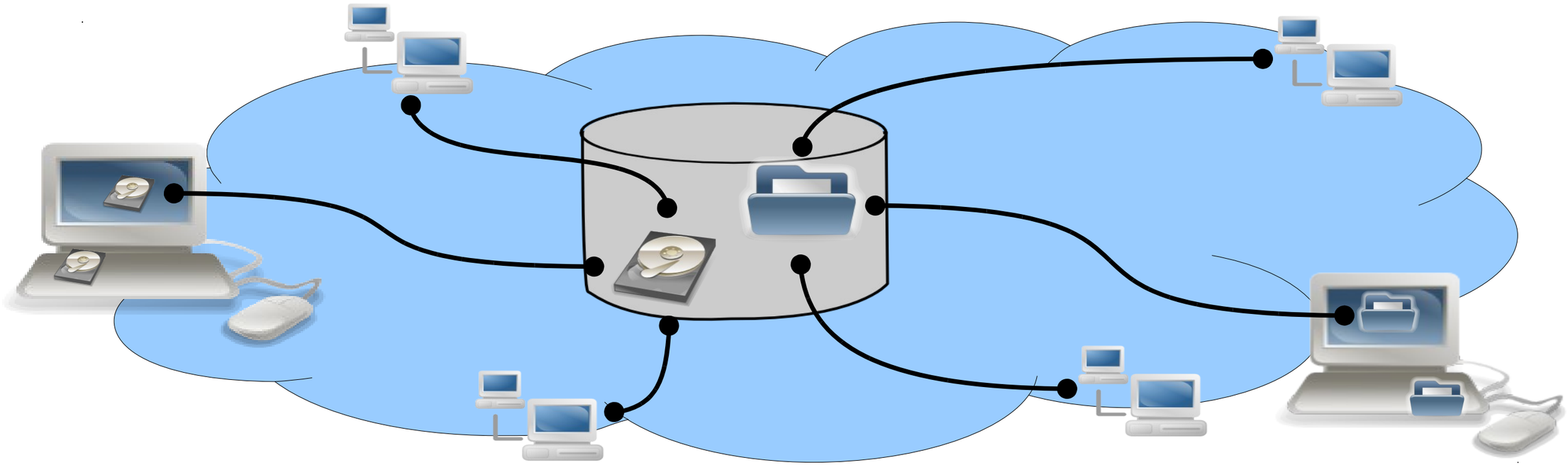
Volume distant

« vu » comme local

Le mode « fichier »

Arborescence distante

« vue » comme locale



ÉCOLE NORMALE SUPÉRIEURE DE LYON

Le cadre de réalisation
COTS & Open Source

- Open Source
  - Pas nécessairement la gratuité : « free as an idea »
  - Reposant sur un socle « normalisé » (IETF, IEEE, ISO)
  - Socle Debian : *« free » but « as a beer & as an idea »*
- COTS
  - Approche « intégrateur »

ÉCOLE **NORMALE**
SUPÉRIEURE
DE **LYON**

COTS : un acronyme (mé)connu ? « Commercial Off-The-Shelf »

Revue « Military & Aerospace Electronics »

COTS : 12 occurrences

Dans « ***Robots in combat missions*** » (page 26)

*« The DOD has mandated the use of **open standards** that **enable interoperability**, such as COM Express, VPX, and PC/104, because of the **cost savings** they deliver. »*

*« Commercial off-the-shelf (**COTS**) components also **allow** manufacturers to improve time-to-market, getting **new technologies deployed** in the field **faster**. »*

*« **Open-standard COTS** components offer even greater advantages, with their modularity **allowing designers** to **develop** more advanced systems **without increased project risk**. »*

ÉCOLE **NORMALE**
SUPÉRIEURE
DE **LYON**

L'approche « Intégration »

- COTS : quatre lettres très en vogue dans l'industrie
  - En développement : « Component Off-The-Shelf »
  - En production : « Commercial Off-The-Shelf »
- « *Off-The-Shelf* » : d'autres...
  - Ont déjà réfléchi à la question
  - Ont trouvé une solution adaptée à un problème
  - Produisent cette solution dans un contexte industriel
  - Assurent le maintien en condition opérationnelle
- « *Component* » ou « *Commercial* »
  - Je prends le « risque » de leur faire « confiance »...

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Dans la forêt des propositions, deux + trois souches retenues

- 2 approches « mode bloc » :
  - iSCSI : le SCSI sur le Réseau IP
  - AoE : Ata Over Ethernet
- 3 approches « mode fichier » :
  - GlusterFS : *le C se transforme en G (comme pour PU)*
  - XtreemFS : *le systèmes de fichiers des grilles*
  - CephFS : *la pieuvre du noyau Linux*

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Mise en œuvre : AoE
La simplicité à l'état pur

- Côté « serveur » ou « cible » :
  - `vblade-persist setup 145 0 eth0 /dev/sda`
    - Active le disque sda sur eth0 avec l'ID 145.0
  - `vblade-persist auto 145 0`
    - Active le lancement automatique de l'ID 145.0
  - `vblade-persist mac 145 0 0 add 00:29:97:92:45:80`
    - Active le filtrage de la machine d'adresse 00:29:97:92:45:80
- Côté « client »
  - `Aoe-discover :` Découverte des périphériques AoE
  - `Aoe-stat :` Visualisation des périphériques AoE détectés
- Ça marche : on découvre par un `dmesg`
  - `aoe: 00093d1367b2 e117.0 v4014 has 286749488 sectors`

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Mise en oeuvre : AoE
Et après ?

- Liste des périphériques : un `aoe-stat` donne :

```
e117.0      146.815GB cluster up
e118.0      146.815GB cluster up
...
e123.0      146.815GB cluster up
e124.0      146.815GB cluster up
```

- Formation d'une grappe RAID6 sur les 8 disques

```
mdadm --create /dev/md5 -binternal --level=6 --force
  --raid-devices=8 /dev/etherd/e???.0
```

- Après, ce que vous voulez dessus :
  - un LVM, un système de fichiers, un iSCSI, un AoE, un GFS, etc...

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Mise en œuvre : iSCSI
Plus compliqué, mais universel

- Côté « serveur » ou « cible » : `iscsitarget`
  - Compilation d'un module noyau (à installer)
  - Configuration : `/etc/iet/ietd.conf`

```
IncomingUser discoveryname DiscJRESPWD

OutgoingUser discoveryname DiscJRESPWD

Target iqn.2008-11.fr.ens-lyon.<MyHostName>:sda
        IncomingUser inputname JRESInPWD
        OutgoingUser outputname JRESOutPWD
        Lun 0 Path=/dev/sda,Type=fileio
```

  - Lancement du démon `iscsitarget`

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Mise en œuvre : iSCSI
Plus compliqué, mais universel

- Côté client : `open-iscsi`
  - Configuration du « découvreur » : `/etc/iscsi/iscsid.conf`

```
discovery.sendtargets.auth.authmethod = CHAP
discovery.sendtargets.auth.username = discoveryname
discovery.sendtargets.auth.password = DiscExpoPWD
```

  - Lancement du « découvreur » sur le noeud X
    - `iscsitarget -m discovery -t st -p X`

```
10.11.12.125:3260,1 iqn.2008-11.fr.ens-lyon.X:sda
172.16.20.125:3260,1 iqn.2008-11.fr.ens-lyon.X:sda
```

  - Accrochage de la « cible » X :
    - `IQN=iqn.2008-11.fr.ens-lyon.X:sda`
    - `iscsistart -a X -i X -t $IQN -g 1 -u inputname -w JRESInPWD`
  - Le disque iSCSI distant apparaît comme un disque SCSI

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Mise en oeuvre : iSCSI
Et après ?

- Liste des périphériques par une recherche dans les logs :
  - `dmesg | grep "Attached SCSI disk"`

```
[39131.733952] sd 6:0:0:0: [sdc] Attached SCSI disk
[39134.032564] sd 7:0:0:0: [sdd] Attached SCSI disk
...
[39144.518260] sd 12:0:0:0: [sdi] Attached SCSI disk
[39146.801402] sd 13:0:0:0: [sdj] Attached SCSI disk
```

- Formation d'une grappe RAID6 sur les 8 disques
  - `mdadm --create /dev/md4 -binternal --level=6 --force --raid-devices=8 /dev/sd[c-j]`
- Après, ce que vous voulez dessus :
  - Un LVM, un système de fichiers, un partage iSCSI ou AoE, un GFS, ...

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Mise en œuvre : GlusterFS
Simple et complet

- Côté « serveurs » ou « cible »
  - Préparer un espace de partage :
    - par exemple `/media/sda`
  - Installation de GlusterFS : la dernière est la 3.2
  - Lancement du démon GlusterFS
  - Et c'est tout...

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Mise en œuvre : GlusterFS Simple et complet

- Côté « client »
  - Installation de GlusterFS : la dernière est la 3.2
  - Lancement du démon GlusterFS
  - Sollicitation des « cibles » : par exemple, la cible X1
    - gluster peer probe X1
  - Création du volume : par exemple, un RAID1
    - `gluster volume create raid replica 2 transport tcp X1:/media/sda X2:/media/sda X3:/media/sda X4:/media/sda X5:/media/sda X6:/media/sda X7:/media/sda X8:/media/sda`
  - Démarrage du volume créé :
    - `gluster volume start raid`
  - Montage du volume démarré
    - `mount -t glusterfs localhost:raid /media/glusterfs.raid`

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Mise en œuvre : XtreemFS
Venant de la grille, basé sur Java

- Côté « serveurs » ou « OSD »
  - Installation de XtreemFS : la version 1.3
    - Paquets xtreemfs-tools, xtreemfs-server, xtreemfs-backend
  - Configuration d'un point de partage :
  - Configuration de l'OSD :
    - définition dans osdconfig.properties
      - D'un UUID unique
      - De l'adresse du maître MRC
      - Du point de partage
      - D'une « passphrase »
  - Démarrage de l'OSD

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Mise en œuvre : XtreemFS
Venant de la grille, basé sur Java

- Côté « maître » et « client » (la même machine!)
  - Installation de XtreemFS : la version 1.3
    - Paquets xtreemfs-tools, xtreemfs-server, xtreemfs-backend
  - Définition du MRC et du DIR
    - Définition de la passphrase, du chiffrement SSL, etc.
  - Démarrage des démons MRC et DIR
  - Création d'un volume
    - `mkfs.xtreemfs -w 2 --chown-non-root M/xtreemfs`
  - Montage du volume
    - `mount.xtreemfs M/xtreemfs /media/xtreemfs`

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Mise en œuvre de CephFS
Une intégration au noyau Linux

- Installation des paquets
- Configuration initiale des « cibles »
  - Formatage d'une partition au format ext3 (pas BtrFS...)
  - Montage avec les attributs étendus : `user_xattr`
- Configuration des « cibles » à partir du « maître »
  - Définition globale dans un fichier unique
    - MON : chaque serveur moniteur
    - MDS : chaque serveur de méta-données
    - OSD : chaque volume distant déclaré

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Mise en œuvre : CephFS
Une intégration au noyau Linux

- Propagation de la configuration
  - `mkcephfs -a -c v22z.conf -k v22z.keyring`
- Démarrage de l'architecture
  - `/etc/init.d/ceph -a start`
- Montage du volume
  - `ceph-fuse -o big_writes /media/cephfs/`
- Et c'est tout...

ÉCOLE **NORMALE**
SUPÉRIEURE
DE **LYON**

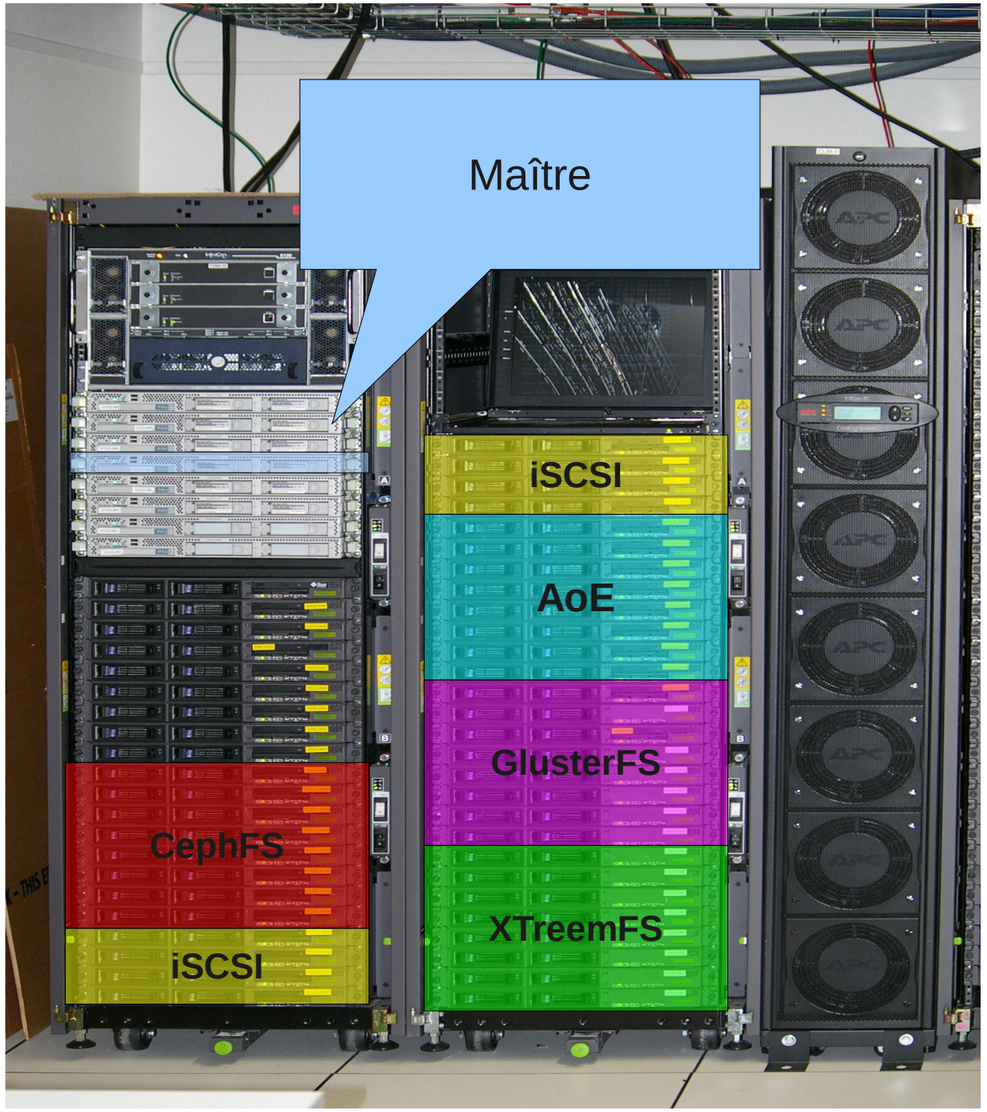
Tableau récapitulatif

Mode	**Mode « bloc »**		**Mode « fichier »**		
Type	**AoE**	**iSCSI**	**GlusterFS**	**XtreemFS**	**CephFS**
Socle	Néant		Ni BtrFS Ni Nilfs		Ext3/Ext4
Client	Tous !		POSIX ACL		
Réplication	MDADM, ZFS, BtrFS		*Replicated*	Totale ou locale	Locale
Snapshot	LVM, ZFS, BtrFS, NilFS		Non	Non	Oui
Disponibilité	Forte !		Bonne	Expérimentale	Expérimentale
Robustesse	MDADM, ZFS		Duplication de serveurs de méta-données		
Intégrité	Non	SCSI !	Intégrée	Intégrée	Intégrée
Confidentialité	Dmcrypt		Aucune !	SSL clé entre noeuds	Clé entre noeuds

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Banc de test
La récupération de machines

- « targets » ou « OSDs »
  - 48 nœuds Sun v22z
  - Composés de
    - bi-cœurs bi-processeurs
    - 8 ou 16 Go de RAM
    - 2 disques SCA de 146 Go
    - 1 connexion GE
    - 1 connexion IB SDR
- « master » ou « client » :
  - 1 nœud Sun x2100



ÉCOLE NORMALE SUPÉRIEURE DE LYON

Banc de test
A l'origine, un *diskless* par tmpfs

- A l'origine, aucun système sur le disque interne.
- Chaque nœud n'a aucun système embarqué. Il :
  - Récupère une IP par DHCP
  - Démarre par PXE et récupère un PXElinux
  - Charge un volume distant NFSROOT en RO
  - Démarre le système
  - Crée dans /etc/fstab autant de TMPfs que nécessaire :
    - /tmp ,
    - /var/run, ...
- A chaque reboot une machine « vierge »

ÉCOLE NORMALE SUPÉRIEURE DE LYON

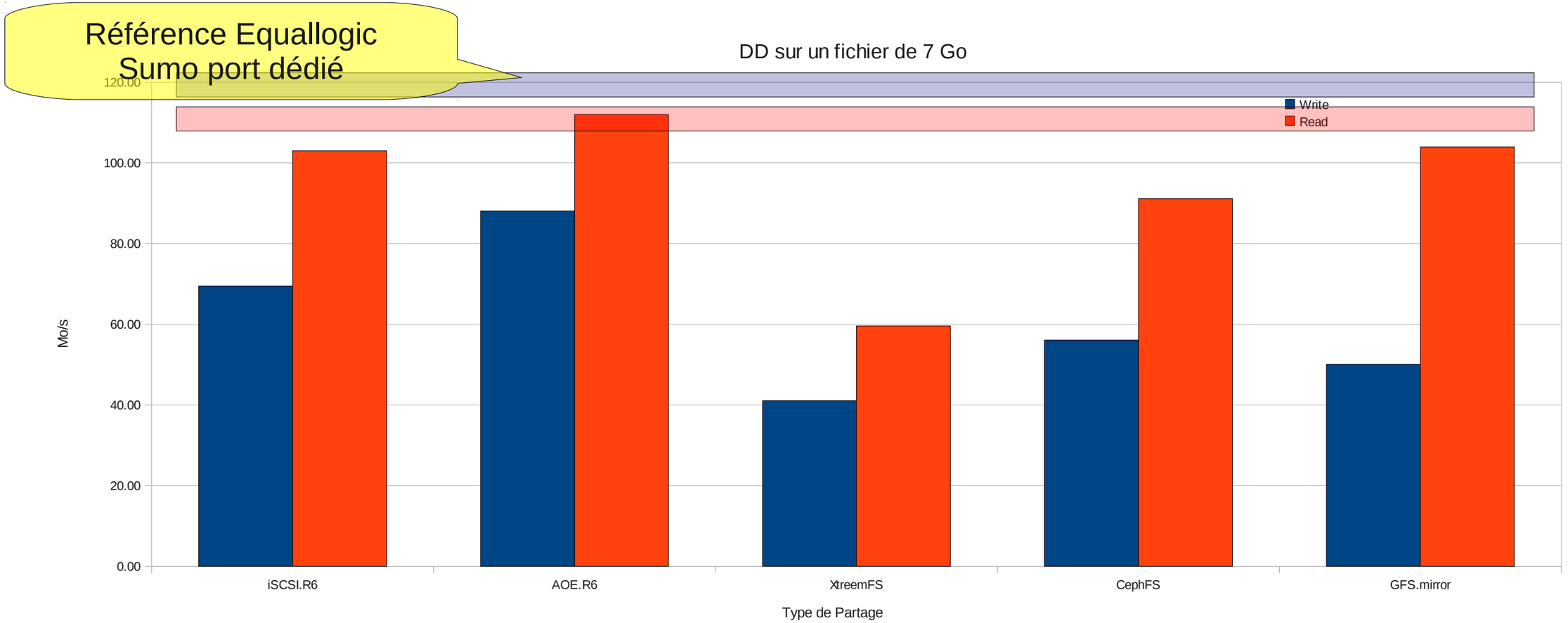
Banc de test
Une nécessité : la persistance

- A l'origine, aucun système sur le disque interne.
- Chaque nœud n'a aucun système embarqué. Il :
  - Récupère une IP par DHCP
  - Démarre par PXE et récupère un PXElinux
  - Charge un volume distant NFSROOT en RO
  - Superpose ce volume à un volume iSCSI en RW
  - Démarre le système de base
  - Configure certains périphériques annexes (IB)
- La persistance assurée par le volume iSCSI
  - Machine « offrant » plus de 80 partages iSCSI

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Test Dédé
(pour « gratter » les disques ☺)

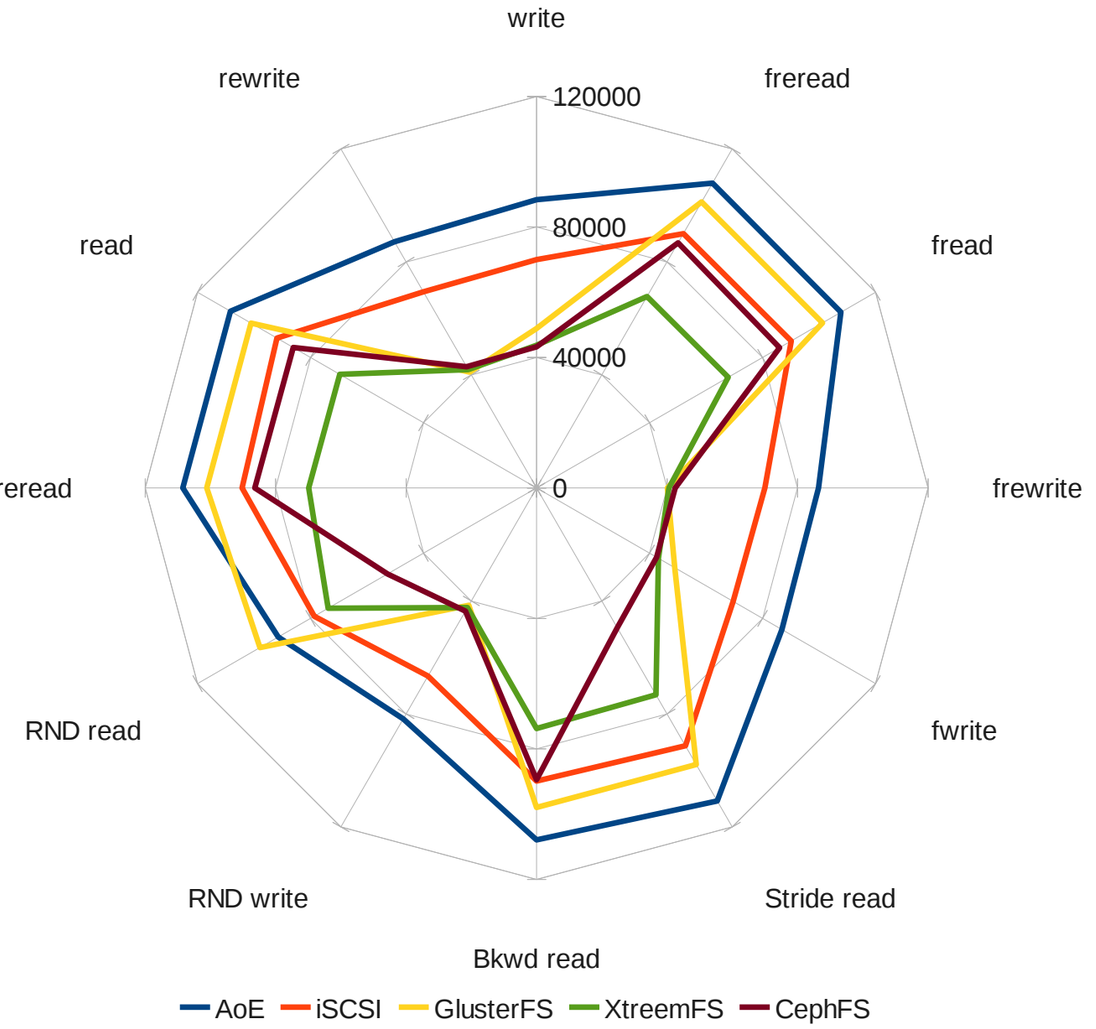
- Ecriture : `dd if=/dev/zero of=/media/distonet/vrac bs=2048k count=3330`
- Lecture : `dd if=/media/distonet/vrac of=/dev/null bs=2048k count=3330`



ÉCOLE NORMALE SUPÉRIEURE DE LYON

Test Iozone3

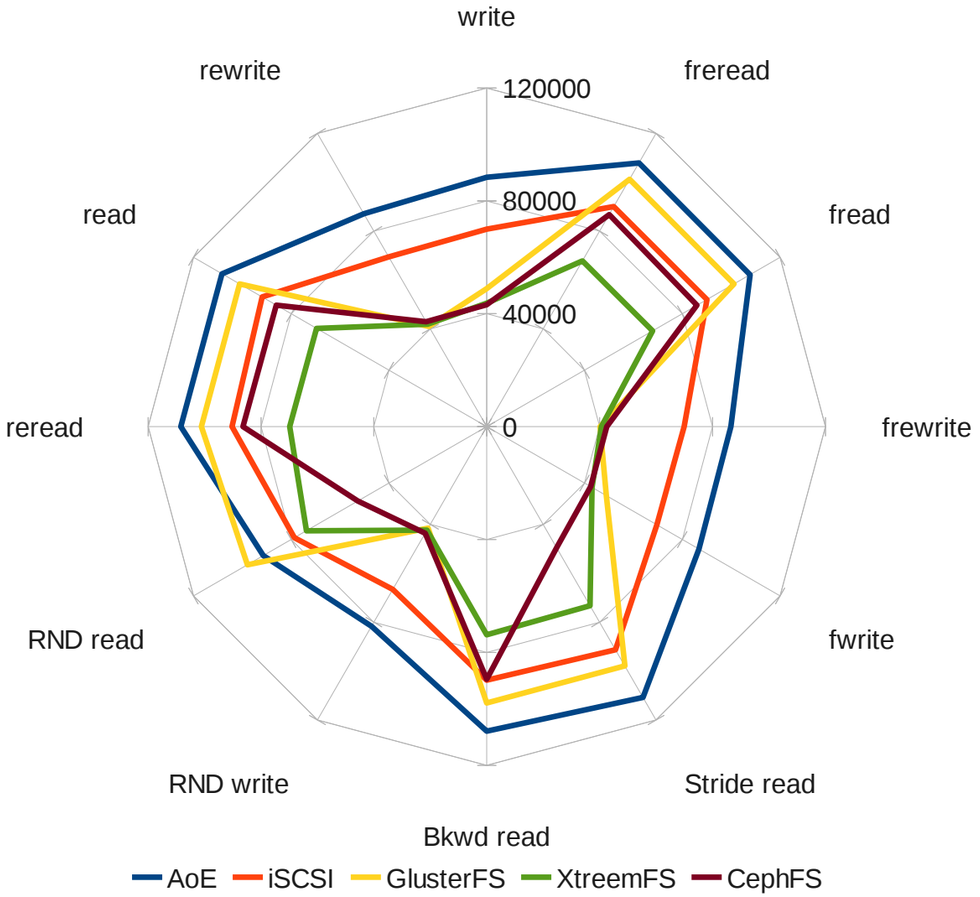
- Meilleur : AoE
- « mode bloc »
  - Équilibrés sur I/O
- « mode fichier »
  - Efficaces en Input
  - GlusterFS
    - Efficace
  - CephFS
    - A suivre...
  - XtreemFS
    - Pas si mal !

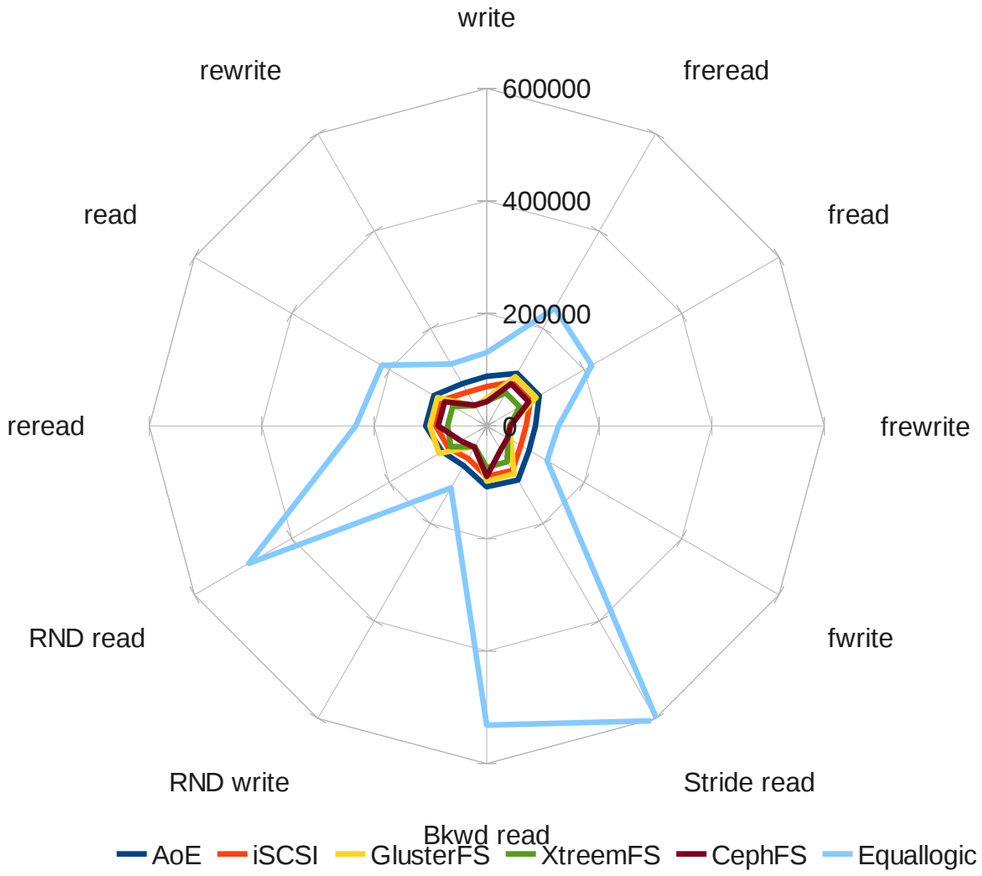


ÉCOLE NORMALE SUPÉRIEURE DE LYON

Test Iozone3 en comparaison avec un SAN...

Avec un Equallogic Sumo, quand même...





ÉCOLE NORMALE SUPÉRIEURE DE LYON

Et sur les postes de travail ?

- Au CBP, parmi les demandes récurrentes :
  - « Environnement de travail » personnalisé :
    - Pour l'utilisation de logiciels spécifiques
    - Pour exploiter les ressources de son poste personnel
    - Pour offrir aux étudiants/chercheurs un cadre générique
  - Solution : une machine virtuelle comparable aux autres

- VirtualBox Open Source « *... it is also the only professional solution that is freely available as Open Source Software under the terms of the GNU General Public License (GPL) version 2.* »

- Système Debian Linux amd64 ou i386 Squeeze :
  - Tous les logiciels science-* installés plus les logiciels spécifiques...

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Machine virtuelle comme socle : rapporté au problème précédent !

- Pourquoi ?
  - Solution (encore) sous contrôle
  - Liaison entre système hôte et VM
  - Espace de disque « à discrétion »
  - Administration à distance
- Comment ?
  - Autonome : système sur espace disque
  - Distante : démarrage par PXE NFSROOT ou iSCSI
- Liaison ?
  - Réseau direct
  - Réseau privé indirect (via passerelle OpenVPN)

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Les futurs travaux

- Côté solutions :
  - En mode « bloc » :
    - ZFS on Linux pour les disques AoE ou iSCSI
  - En mode « fichier » :
    - Suivre CephFS
    - Expérimenter les autres :
      - OpenStack : solution choisie par Ubuntu
      - Coda : pourquoi si peu connu ? Intérêt pour la mobilité ?
      - Hadoop (mais Java seulement Sun)
      - MooseFS, Gfarm, ...
- Côté déploiement :
  - Cluster, Salle de cours, machines spécifiques, et après ?

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Envie de joindre le projet ?

- Une adresse :
  - http://forge.cbp.ens-lyon.fr/projects/distonet
- Deux interlocuteurs :
  - Emmanuel Quémener : côté « expérimental »
  - Loïs Taulelle : côté « production future »