iSCSI : de l'expérimentation à la mise en production

Bernard DEBORD
UFR/IMA Université Joseph Fourier - BP 53 -
38 041 GRENOBLE cedex 9
bernard.debord@imag.fr

Sigrun FREDENUCCI
DSI Grenoble Universités
B.P. 53 38041 GRENOBLE CEDEX
Sigrun.Fredenucci@grenet.fr

Didier MATHIAN
DSI Grenoble Universités
B.P. 53 38041 GRENOBLE CEDEX
Didier.Mathian@grenet.fr

Pascal PRALY
CRI Université Pierre Mendès France
151, avenue des Universités
38 400 Saint-Martin d'Hères
Pascal.Praly@upmf-grenoble.fr

Résumé

Cet article rapporte l'expérimentation des technologies iSCSI et leur application pour la mise en œuvre de nouvelles stratégies de sauvegardes à Grenoble Universités.

Nos expérimentations ont porté sur les aspects de performance, de robustesse, de sécurité en faisant varier les « initiateurs » (différents OS et différents pilotes iSCSI), les réseaux IP et les « cibles » (disques accessibles via une passerelle IP-FC ou une passerelle IP-SCSI).

Les résultats encourageants de ces expérimentations et la promesse de réseaux IP toujours plus rapides nous ont décidé à franchir le pas vers une mise en œuvre.

L'actuel projet a comme objectif opérationnel la mise en place d'une politique de sauvegardes à distance pour les laboratoires de l'Université Joseph Fourier qui puisse être étendue à d'autres composantes de Grenoble Universités.

Mots clefs

iSCSI, SAN, FC, Sauvegardes extra muros

1 Introduction

Grenoble Universités présente une expérience importante dans le domaine du Storage Area Network (SAN). La recherche de performance et de robustesse nous a conduit à choisir la technologie Fibre Channel (FC) pour la constitution de nos SAN ; elle a donné entière satisfaction et se justifie pleinement pour des applications critiques (comme les clusters APOGEE). Cependant le coût de la mise en place ne permet pas de généraliser son usage : la multiplication des îlots de stockage est rapidement limitée ; la constitution d'îlots multi-sites demande une infrastructure de fibres dédiées trop importante.

Voila pourquoi nous avons tourné nos efforts vers l'expérimentation de la technologie iSCSI qui nous parait complémentaire à la technologie FC. En effet, il s'agit toujours de rendre accessibles des espaces disques à des serveurs distants mais cette fois-ci en profitant de l'infrastructure IP existante. Les coûts de mise en œuvre sont donc réduits moyennant une baisse des performances.

2 Le protocole iSCSI

Le protocole Internet SCSI (iSCSI) est, comme son nom l'indique, une encapsulation du protocole SCSI dans TCP/IP ; il permet d'accéder en mode bloc, par le réseau, à des unités de stockage (disques, robots, lecteurs de disques optiques...). De ce fait il peut être utilisé par des applications de consolidation de stockage, de clusters de serveurs, de virtualisation de serveurs, de sauvegarde sur bande ou encore des plans de reprise d'activité (PRA).

Le point fort de ce protocole est d'utiliser l'infrastructure TCP/IP existante et non pas une infrastructure dédiée comme le protocole FC tout en fournissant des fonctionnalités analogues.

Son point faible par rapport au protocole FC est une moindre performance, et ceci même sur des réseaux dédiés.

Si on compare le protocole iSCSI à des protocoles comme NFS ou SAMBA, iSCSI se situe au niveau bloc (comme avec le FC, on donne accès à des disques que l'on formate et utilise ensuite à sa guise) alors que NFS ou SAMBA se situent au niveau fichier (on donne accès à des répertoires et des fichiers avec leurs permissions). Les performances sont meilleures en iSCSI car elles se situent à un niveau plus bas dans les couches logicielles ce qui le rend intéressant pour des applications de sauvegardes ou de PRA qui nécessitent de gros transferts de données.

2.1 Architecture

L'architecture iSCSI est basée sur un modèle client serveur.

Le client ou initiateur est actif, par exemple une machine qui demande une lecture ou une écriture de blocs.

Le serveur ou cible est passif, par exemple une unité de stockage, sur demande de l'initiateur lui fournit un certain nombre de blocs de données.

Le serveur contient des Logical Unit (désignées par leur Logical Unit Number : LUN) qui traitent les commandes des initiateurs. Les commandes sont contenues dans des Command Descriptor Block (CDB).

Chaque commande est traitée dans une seule connection TCP.

Le but principal du iSCSI est d'encapsuler et d'acheminer de manière sûre les CDB entre initiateurs et cibles sur TCP/IP.

2.2 Terminologie

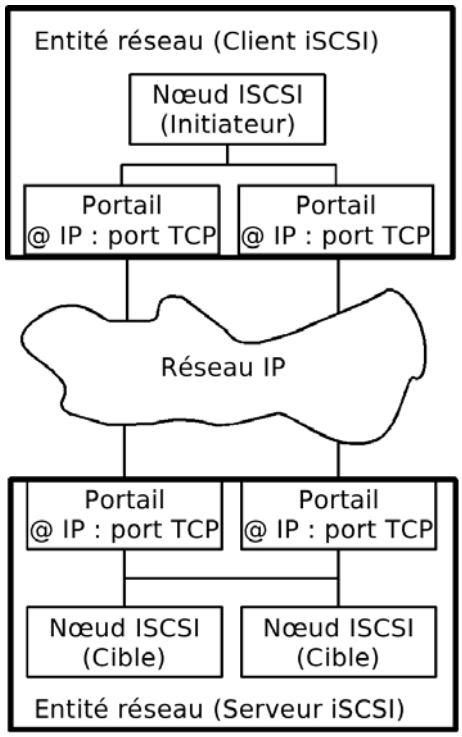


Figure 1 - *Terminologie iSCSI*

Reprenons la terminologie telle que définie dans la rfc 3720 [1],voir Figure 1.

Une entité réseau représente un client iSCSI (e.g. une machine) ou un serveur iSCSI (e.g. une passerelle IP/FC) accessible depuis le réseau IP.

Une entité réseau doit avoir un ou plusieurs portails qui peuvent être utilisés par un nœud iSCSI pour accéder au réseau.

Un nœud iSCSI est l'analogue d'un périphérique SCSI (initiateur ou cible).

2.3 Adressage et nommage

Un nœud iSCSI est désigné par un nom indépendant de sa situation.

Les noms respectent l'un des deux formats :

1. le format iqn (iSCSI qualifier name)

   iqn.fr.ujf-grenoble.ela-1400-ufrima

2. le format eui-64 (Extended Unique Identifier définit par l'IEEE [2])

   eui.200000D02300FFFF

   eui.wwn FC soit 64 bits en hexa

A chaque nom iSCSI correspond un nom standard <domain-name>[:port] où <domain-name> est soit une adresse IP soit un FQN (Fully Qualified Name). Si aucun port n'est spécifié, le port par défaut de l'IANA (Internet Assigned Numbers Authority) 3260 est utilisé.

Le protocole Internet Storage Name Service (iSNS) a pour but de faciliter la découverte des cibles aux initiateurs. Dans son principe, un initiateur interroge un serveur iSNS pour connaître les adresses IP des cibles. Nous n'avons pas testé de serveurs iSNS.

2.4 Session iSCSI

Le niveau le plus élevé d'une communication iSCSI est une session qui est ouverte entre un initiateur et une cible.

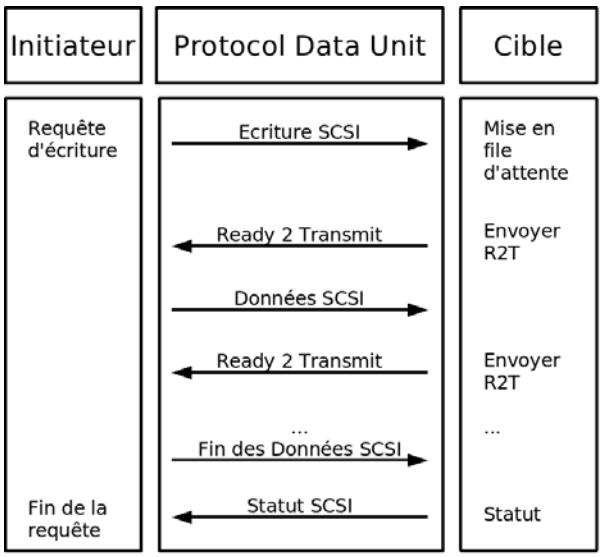
Deux types de sessions sont possibles, une session de découverte utilisée par l'initiateur pour découvrir les cibles disponibles et les sessions « normales » qui permettent les transferts de données.

Une session iSCSI se décompose en trois phases : le login, les transferts et le logout.

La phase de login permet d'authentifier le client auprès du serveur puis de négocier les paramètres de communication (time out, taille paquets...).

Les CDB, les statuts et les données sont acheminées via des Protocol Data Unit (PDU). Les transferts se font au rythme des vidages des tampons d'émission et de réception, voir Figure 2.

Figure 2 - *Exemple d'écriture iSCSI [3]*



Le logout ferme la ou les sessions TCP (voir Figure 3).

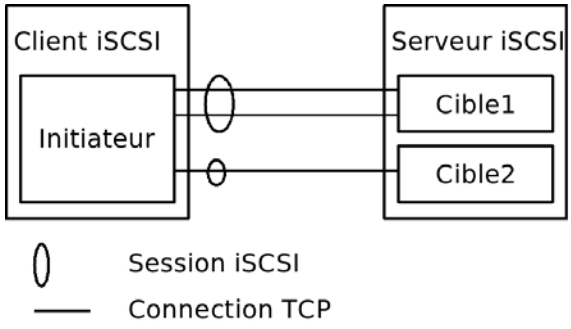


Figure 3 - *Session iSCSI [4]*

3 L'infrastructure de test

Notre infrastructure de test se compose d'un client (serveur DELL 1850, carte réseau intégrée) sous Windows ou Linux, d'un switch 1000Mb/s (ou 100Mb/s) et des serveurs iSCSI suivants :

1. Eclipse 1620 (constructeur McData) + baie FC ELA1400

   L'eclipse est un switch d'interconnexion iSCSI-FC muni de deux ports IP et de deux ports FC, la baie ELA est directement connectée sur un port FC, sur l'autre port nous avons connecté un SAN en production via un switch FC Brocade.

2. IPBridge 1500 (constructeur ATTO) + disques SCSI

   L'IPBridge est un pont d'interconnexion iSCSI-SCSI muni d'un port IP et d'un connecteur SCSI. Dans notre cas, nous avons connecté un disque SCSI, mais nous avons aussi testé avec succès l'accès à un robot de sauvegarde SUN StorageTek L25 sur la même chaîne SCSI.

3. FAS 250 (constructeur NETAPP – disques FC)

   Cette baie a deux ports IP et parle nativement iSCSI.

4. Snap Server 650 (constructeur ADAPTEC – disques SATA)

   Ce Network Attached Storage (NAS) a deux ports IP et parle nativement iSCSI.

3.1 Configuration

Avant de pouvoir utiliser notre infrastructure, il reste à configurer la partie client (par l'intermédiaire du pilote iSCSI) et la partie serveur.

Le pilote iSCSI se combine avec la pile TCP/IP, le pilote réseau et la carte réseau pour fournir le même service qu'un pilote de carte SCSI (voir Figure 4).

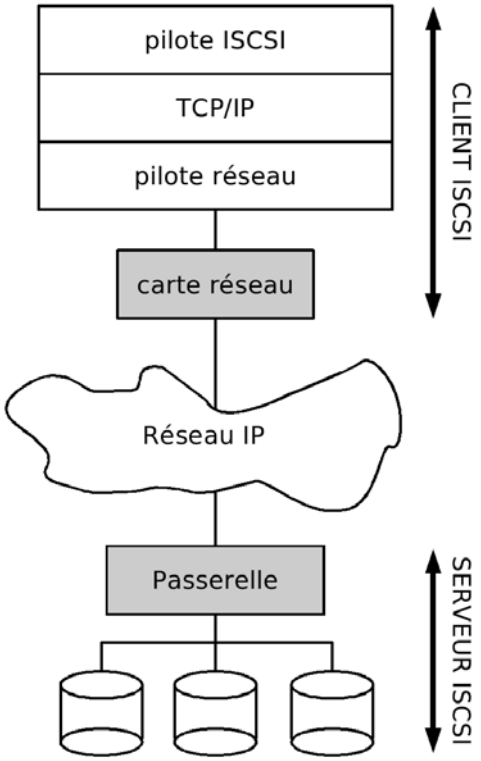


Figure 4 - *Pilote iSCSI*

3.1.1 Client iSCSI linux

Sous linux, le pilote utilisé est open-iscsi.

Après avoir positionné le nom de l'initiateur dans le fichier /etc/iscsi/initiatorname.iscsi, l'utilitaire iscsiadm permet de réaliser les opérations de base.

Découverte des cibles iscsi :

```
# iscsiadm -m discovery -t sendtargets -p 130.190.225.110:3260
130.190.225.110:3260,-1 eui.210000d02380ffff
```

Pour voir la liste des cibles découvertes :

```
# iscsiadm -m discovery
```

```
130.190.225.110:3260 via sendtargets
```

Login ISCSI :

```
# iscsiadm -m node -T eui.210000d02380ffff -p
    130.190.225.110:3260 -l
```

Pour voir les noeuds sur lesquels on est connecté :

```
# iscsiadm -m node
130.190.225.110:3260,-1 eui.210000d02380ffff
```

Les disques sont alors visibles comme de nouveau device ; un extrait du dmesg donne par exemple :

```
scsi2 : iSCSI Initiator over TCP/IP
  Vendor: AXUS      Model: ELA-1400          Rev: 0316
  Type:   Direct-Access                      ANSI SCSI revision: 04
SCSI device sdb: 40960000 512-byte hdwr sectors (20972 MB)
sdb: Write Protect is off
sdb: Mode Sense: 87 00 00 08
SCSI device sdb: drive cache: write through
SCSI device sdb: 40960000 512-byte hdwr sectors (20972 MB)
sdb: Write Protect is off
sdb: Mode Sense: 87 00 00 08
SCSI device sdb: drive cache: write through
 sdb: sdb1 sdb2
sd 2:0:0:0: Attached scsi disk sdb
sd 2:0:0:0: Attached scsi generic sg1 type 0
```

Ce nouveau device /dev/sdb est directement utilisable comme un disque local par des commandes comme fdisk, mkfs...

C'est avec le mécanisme de LUN masking de la baie FC que seul le LUN sdb est visible et accessible pour ce client iSCSI.

Logout ISCSI :

```
# iscsiadm -m node -T eui.210000d02380ffff -p
    130.190.225.110:3260 -u
```

3.1.2 Client iSCSI Windows

Sous windows (VISTA, XP et Win2003 server), nous avons utilisé le pilote fournit par Microsoft « iSCSI Software Initiator 2.x ».

Les Figure 5 et Figure 6 illustrent la phase de découverte des cibles ; la Figure 7 la phase de connection proprement dite.

Dans la Figure 5, le portail cible est un des deux ports IP de l'Eclipse 1620, aucun service iSNS n'est mis en œuvre.

Dans cette phase, le client iSCSI tente une première connection à l'Eclipse. Cette connection lui est refusée car, pour des raisons de sécurité, un client iSCSI inconnu n'a accès à aucune ressource. Il faut une intervention manuelle au niveau de l'Eclipse pour autoriser ce nouveau client à accéder à des LUNs.

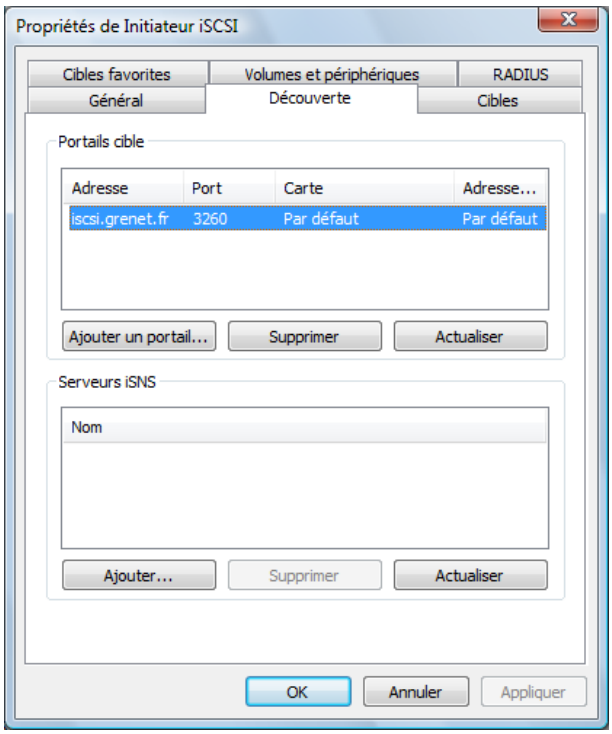


Figure 5 - *Saisie de l'adresse IP du portal serveur*

Une fois cette opération effectuée, des cibles sont maintenant visibles. Dans la Figure 6 on constate que le login iSCSI a permis la découverte de deux cibles.

Le LUN eui.210000d02330ffff est celui déjà utilisé dans le paragraphe précédent pour la connection d'un client iSCSI Linux.

Pour faciliter nos tests, nous avions laissé un LUN ouvert à tous les initiateurs iSCSI et nous avions créé deux partitions l'une que nous avons formaté en ext3 lors des tests linux, et l'autre que nous avions réservé pour les tests windows et que nous avons formaté en NTFS.

Le LUN eui.500601683021a4bb, quant à lui, est une ressource disque sur un SAN en production. Nous l'avons utilisé pour effectuer une sauvegarde journalière de fichiers sous XP à travers un réseau de campus.

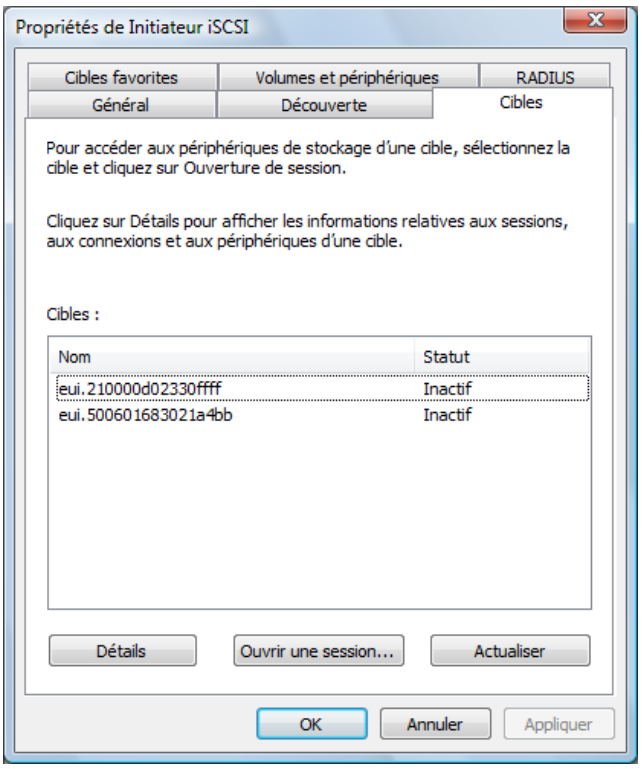


Figure 6 - *Liste des cibles découvertes*

Dans la Figure 7, remarquez la case à cocher « Restaurer automatiquement cette connection au démarrage de l'ordinateur » qui permet de disposer en permanence du ou des périphériques via le portail iSCSI, comme n'importe quel disque interne.

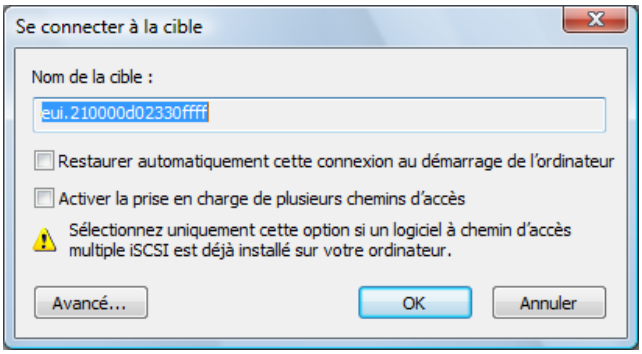


Figure 7 - *Connexion à une cible découverte*

Une fois connecté, les disques sont visibles par le gestionnaire de disque de windows comme n'importe quel disque local.

Il ne reste qu'à formater et à utiliser !

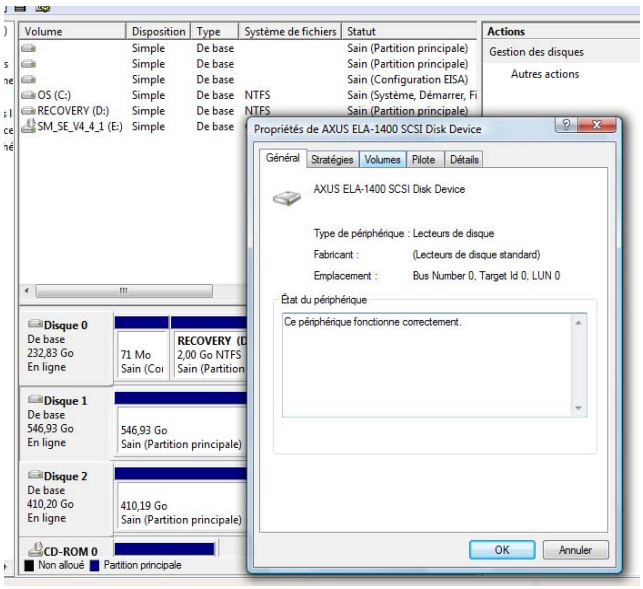


Figure 8 - *Disque1 et Disque2 apparaissent dans le gestionnaire des disques*

3.1.3 *Serveur iSCSI Eclipse 1620*

Le switch d'interconnexion McData Eclipse 1620 rend accessible l'ensemble des données d'un SAN via iSCSI.

Il se configure via une interface d'administration qui permet de faire du zoning par port.

Dans la Figure 9, le pc-de-mathian a le droit de voir :

1. le SAN connecté au port 1 via le switch Brocade
2. la Baie ELA directement connecté au port 2.

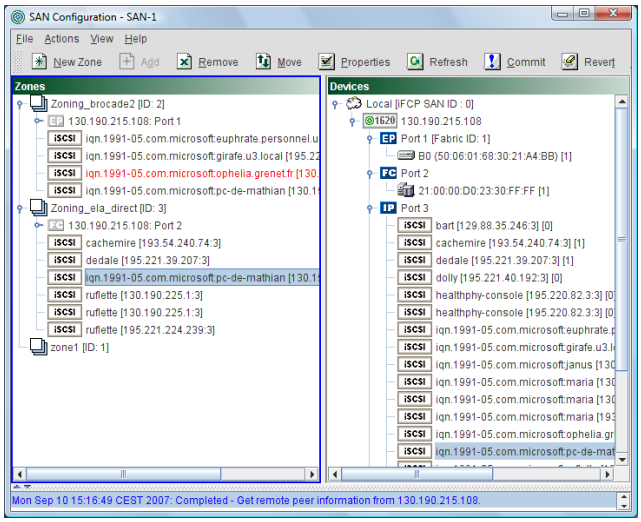


Figure 9 - Configuration eclipse 1620

3.1.4 *Serveur iSCSI IPBridge*

Le serveur iSCSI IPBridge 1500 rend accessibles en iSCSI des périphériques SCSI. Nous avons testé l'accès à des disques SCSI ainsi qu'un robot de sauvegardes L25. La sécurité au niveau du boîtier peut utiliser le Challenge Handshake Authentication Protocol (CHAP) ou être déléguée au niveau d'un serveur iSNS.

3.1.5 Serveurs iSCSI FAS 250 et Snap Server 650

Les NetApp FAS 250 et Adaptec Snap Server 650 sont des NAS qui offrent le protocole iSCSI en plus des protocoles NAS classiques : CIFS, NFS, FTP, HTTP... Il faut créer des LUNs sur ces NAS pour pouvoir y accéder par iSCSI. Ceci se fait par l'interface Web d'administration des NAS.

3.2 Sécurité

La sécurité d'accès aux données peut être réalisée à plusieurs niveaux.

Le protocole iSCSI prévoit dans sa phase de login la possibilité d'effectuer une authentification (Kerberos, CHAP...) auprès du serveur iSCSI. Sur le NetApp par exemple, l'authentification par CHAP peut être activée ; dans ce cas, pour chaque initiateur, on peut définir si on utilise ou non cette méthode d'authentification.

Une fois connecté, la passerelle peut avoir ses propres moyens d'assurer la sécurité. Sur l'Eclipse par exemple, il faut lier explicitement l'initiateur à un port FC.

Dans le cas d'un accès à une baie de disques FC, il est encore possible de faire du LUN masking, c'est-à-dire de ne permettre l'accès d'un LUN qu'à un certain initiateur.

Le NAS NetApp permet aussi de faire du LUN masking. Pour cela, il faut créer un groupe d'initiateurs iSCSI comprenant un ou plusieurs clients et définir si on fait un « map » du LUN pour ce groupe. Une fois le « map » activé, les clients du groupe ont alors accès au LUN.

3.3 Performances

Un de nos tests correspond à la création d'un fichier de 2Go avec mesure du temps d'exécution et du pourcentage d'utilisation CPU.

Ce qui donne sous Linux :

```
/usr/bin/time -f « CPU %P real %es » sh -c « dd if=/dev/zero
of=testfile bs=64k count=32768; sync »
```

Pour Windows, nous n'avons pas de mesure exacte car nous n'avons pas trouvé de moyen de déclencher la synchronisation des disques (le sync UNIX) après la copie. Nous avons constaté en effet qu'une fois que le système rend la main – on croit que la copie est finie -, l'activité réseau continue ce qui fausse complètement les mesures. Pour avoir néanmoins des approximations nous avons fait des mesures chronomètre en main en surveillant l'activité réseau et l'activité sur le serveur iSCSI. Il semblerait que les performances soient voisines de celles obtenues sous linux.

Dans notre infrastructure 1, sous linux, nous avons mesuré des débits de :

- 93Mb/s sur un switch 100Mb/s
- 414 Mb/s sur un switch 1Gb/s

Dans notre infrastructure 2, sous linux, nous avons mesuré des débits de :

- 93Mb/s sur un switch 100Mb/s
- 148 Mb/s sur un switch 1Gb/s

Dans notre infrastructure 3, sous linux, nous avons mesuré des débits de :

- 240 Mb/s sur un switch 1Gb/s

(En comparaison, nous avons obtenu un débit de 620Mb/s sur une partition en RAID1 locale au serveur linux)

Dans notre infrastructure 4, sous linux, nous avons mesuré des débits de :

- 93 Mb/s sur un switch 1Gb/s.

4 Solutions opérationnelles

Sur un réseau de campus (avec des goulets d'étranglement à 100Mb/s) sur des plates-formes en production nous avons mesuré des performances de :

- 50Mb/s depuis XP ou Win2003
- 40Mb/s depuis Linux.

Ces performances sont suffisantes pour déclencher des sauvegardes de petits volumes (600Mo par exemple pour une sauvegarde journalière de fichiers sous XP).

Pour des volumes plus importants (400Go pour les HOMES des étudiants de l'UFRIMA constitués d'une multitude de petits fichiers) nous avons opté pour une réplication par l'utilitaire rsync qui ne transfère que les deltas entre l'original et la copie. Dans ce cas précis, contre toute attente, le temps de sauvegarde par iSCSI est voisin de celui par FC (sur une infrastructure à 1Gb/s). Il s'avère en effet que le temps de transfert des deltas est négligeable par rapport au temps de calcul des différences !

Dans ces deux cas de figure, nous n'avons pas eu de problème de fiabilité, les sauvegardes ayant lieu chaque jour avec succès.

5 Conclusion et perspectives

La technologie iSCSI est à l'heure actuelle une technologie mature et stable. Elle permet de déporter des unités de stockage loin des serveurs qui les utilisent ce qui la rend particulièrement intéressante pour des applications de virtualisation de serveur et de sauvegarde.

Son avantage principal par rapport au FC est un moindre prix du fait de l'utilisation de l'infrastructure TCP/IP existante (pas de fibres dédiées, pas de switch ni de cartes dédiés).

Malgré tout, les performances sont moindres qu'en FC (Ethernet fonctionne à 1Gb/s contre 4Gb/s pour le Fiber Chanel).

Notre interrogation principale concerne la fiabilité. Nous n'avons pas constaté ce problème pour le moment, mais nous ne sommes pas prêt pour autant à faire reposer une application critique sur des disques distants accessibles par TCP/IP. C'est pourquoi nous avons axé nos mises en production sur de la sauvegarde ou des PRA en attendant d'être convaincu pour des applications plus critiques.

Bibliographie

[1] Internet Small Computer Systems Interface (iSCSI), RFC 3720, http://www.ietf.org/rfc/rfc3720.txt

[2] IEEE, "Guidelines for 64-bit Global Identifier (EUI-64) Registration Authority", http://standards.ieee.org/regauth/oui/tutorials/EUI64.html , mars 1997.

[3] SNIA IP Storage Forum, « iSCSI Technical Overview », http://www.snia.org.

[4] CISCO, « Introduction to iSCSI », white paper, http://www.cisco.com