Retour d'expérience SAN multi site: de problèmes en solutions

Bernard Debord
UFR/IMA Université Joseph Fourier
UFRIMA - BP 53 - 38 041 Grenoble Cedex
bernard.debord@ujf-grenoble.fr

Sigrun Fredenucci
DSI de Grenoble Universités
ex CICG BP 53 38041 Grenoble CEDEX 9
Sigrun.fredenucci@grenet.fr

Didier Mathian
DSI de Grenoble Universités …
ex CICG BP 53 38041 Grenoble CEDEX
didier.mathian@grenet.fr

Résumé

Cet article rapporte la mise en œuvre d'un réseau de stockage (Storage Area Network, SAN) multi site sur le Domaine Universitaire grenoblois.

Ce SAN multi site, basé sur le protocole Fibre Channel (FC), est constitué de trois commutateurs FC et de trois baies FC localisées dans deux bâtiments distants de 200m ; il sert à ce jour à l'hébergement de bases de données Apogée pour test et validation, et à la sauvegarde avec une étape sur disque intermédiaire pour le logiciel de sauvegarde par réseau Netbackup. Suite à ce projet pilote, l'UFR/IMA et la DSI/GU se sont équipées de commutateurs et de baies FC plus évolués lors de l'évolution de leur infrastructure de stockage.

L'actuel projet a comme objectif opérationnel la mise en place d'une politique de sauvegardes permettant de diminuer de façon significative la fenêtre de sauvegardes pour des volumes de données importants (répertoires d'accueil des étudiants, bases de données Oracle) et comme objectif second de jeter les fondements d'un Plan de Reprise d'Activité (PRA) du Système d'Information de Grenoble Universités.

Mots clefs

SAN, HBA, FC, FCP, iFCP, Réseau de stockage, Sauvegarde, Haute disponibilité, Plan de reprise d'activité…

1 Introduction

L'utilisation des technologies de l'information et de la communication (TIC) est devenue indispensable aussi bien à la diffusion et à la progression des connaissances qu'au fonctionnement même du système d'information (SI) des établissements. Une partie toujours plus importante du savoir et du savoir faire de nos universités se trouve sous forme numérique ; ces données sont de jour en jour plus volumineuses et plus critiques.

Les centres de ressources informatiques qui assurent la gestion de ces données sont confrontés à deux demandes antagonistes : d'un coté des volumes de données toujours plus importants et plus sensibles nécessitent des sauvegardes journalières qui prennent de plus en plus de temps ; et d'un autre coté, le temps d'indisponibilité des applications qui manipulent ces données est toujours plus réduit.

De plus, la criticité des données manipulées pose le problème de sinistres éventuels et de la mise en place d'un véritable Plan de reprise d'activité (PRA).

Cet article se propose de parcourir la problématique de la mise en place d'un réseau de stockage dans deux environnements:

- la composante UFR/IMA, site pilote de l'Université Joseph Fourier pour l'utilisation des technologies réseau de stockage (SAN)

- le département Informatique de Gestion de la DSI de Grenoble Université (ex CICG), pour un projet d'évolution de son infrastructure de stockage basée sur le SAN

Dans une première partie, nous présenterons un tour d'horizon des différentes technologies SAN.

Dans une seconde partie, nous exposerons la mise en place du réseau de stockage expérimental multi site.

Dans la troisième partie nous présenterons les solutions opérationnelles pour l'ébauche d'un plan de reprise d'activité mis en œuvre dans les deux sites participants au projet SAN.

Enfin, nous présenterons les résultats et retours d'expériences de notre projet ainsi que ses perspectives d'évolution.

2 Les technologies SAN dans l'environnement Fibre Channel

2.1 Le protocole Fibre Channel

Le protocole Fibre Channel allie le meilleur du stockage et du réseau.

Il a été construit pour dépasser les limitations du protocole SCSI à savoir principalement la longueur de câble (de 30m on passe à 10km) tout en conservant ses points forts (pas de surcharge processeur, grande occupation de la bande passante) et en augmentant les débits.

Le protocole SCSI est une architecture de bus (parallèle) Half-Duplex (l'information ne circule que dans un sens à la fois).

Le protocole FC est une architecture réseau (série) Full-Duplex (l'information circule dans les deux sens en même temps ce qui explique la présence de deux câbles, un pour l'émission l'autre pour la réception, chaque port FC comportant un transmetteur et un récepteur).

On pourrait penser qu'une transmission parallèle est plus rapide qu'une transmission série ; en fait, c'est l'inverse qui se produit. Les fréquences de transfert sur les bus doivent assurer que les données transmises sur chaque câble arrivent dans le même temps d'horloge, alors que pour une transmission série, l'information est transmise aussi rapidement que le média le supporte.

Quelques caractéristiques du protocole FC :

- Il fonctionne en mode bloc : les données sont segmentées en trames de 2 148 octets maximum (cf. Ethernet 1 518 octets ou jumbo frames 4 096 octets); les trames sont groupées en séquences autorisant des transferts de 128Mo.

- Il reprend le vocabulaire du protocole SCSI (Initiateur, Cible, Séquence).

- Les spécifications FC sont disponibles pour les vitesses 1Gb/s 2Gb/s 4Gb/s.

- Les équipements FC sont enfichables à chaud.

2.2 Les 5 niveaux du protocole FC

Le protocole FC est structuré en niveaux comme le modèle OSI sans être cependant identique à celui-ci.

2.2.1 FC-4

Décrit l'interface entre FC et d'autres protocoles de haut niveau (IP, ATM, SCSI…).

2.2.2 FC-3

Fonctionnalités avancées (Common Services) : multicast, compression, chiffrement.

2.2.3 FC-2

Ce niveau décrit les topologies possibles, les types de ports, coupe les données en trames puis les ré assemble après transport, route les données.

2.2.4 FC-1

Ce niveau décrit les règles de transmission, le codage et décodage (8bits vers 10bits), le contrôle d'erreurs, l'accès au média.

2.2.5 FC-0

Ce niveau décrit le media physique (cuivre, fibre, transmetteurs, receveurs, distances, connecteurs…). Par exemple pour le cuivre la distance est limitée à 30m, pour la fibre multimode 250m et pour la fibre monomode 10km avec un répéteur.

Chaque équipement est identifié par un numéro unique le WWN (World Wide Number).

2.3 Les différents types de ports

- N_PORT (Node): permet une connexion point à point vers un autre N_PORT ou vers un F_PORT sur un commutateur

- NL_PORT (NodeLoop): c'est un N_PORT qui peut participer à un « Arbitrated Loop »

- F_PORT (Fabric) : sont utilisés sur un commutateur pour connecter un N_PORT

- FL_PORT (FabricLoop) : sur un commutateur permet de participer à un « Arbitrated Loop »

- E_PORT (Expansion) : permet de connecter un commutateur à un autre commutateur

- G_PORT (Generic) : sur un commutateur se comporte comme un E_PORT, FL_PORT ou F_PORT

2.4 Les différentes topologies de SAN Fibre Channel

Un îlot SAN est un ensemble d'équipements FC reliés entre eux ; pour un îlot trois topologies sont possibles.

2.4.1 Topologie « Point à point »

Il s'agit d'une connexion directe entre deux N_PORTs (dont un initiateur).

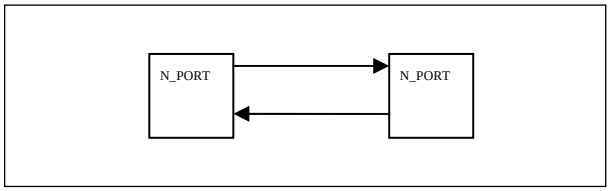

Figure 1- *Point à point*

2.4.2 Topologie en « Arbitrated Loop »

Sur un « Loop » un seul nœud peut envoyer des données à la fois, pour accéder à une ressource un nœud doit gagner un arbitrage (comme en SCSI). Les différents équipements partagent donc la bande passante.

L'adressage des ports se fait sur 8 bits, chaque « Loop » peut contenir 126 nœuds NL_PORT (mais cette limite est théorique car les performances chutent au-delà de 30).

Un « Arbitrated Loop » peut être réalisé de deux manières différentes, en chaînant directement les équipements entre eux ou bien en les connectant sur un hub. Cette dernière topologie est plus robuste car un équipement peut être retiré ou ajouté sans rompre le « Loop » grâce à l'électronique embarquée dans le hub.

Les « Private Arbitrated Loop » ne voient que des équipements du « Loop » ; les « Public Arbitrated Loop » peuvent être reliés à un « Fabric » via un port FL_PORT.

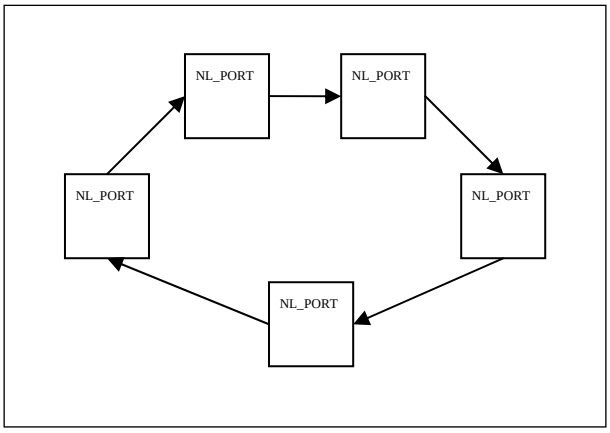

Figure 2-*Arbitrated Loop*

2.4.3 Topologie de type « Fabric »

Dans le mode « Fabric » les trames peuvent être acheminées directement d'un port à un autre, les commutateurs permettent l'agrégation de bande passante.

L'adressage des ports se fait sur 24 bits et permet de connecter plus de 16 millions d'équipements.

Lors de la connexion, le service de login du « Fabric », sur réception du WWN, retourne une adresse physique, le FCID, qui sera utilisée pour le routage des trames.

Le FCID est divisée en trois champs, un pour le domaine (un commutateur), un pour l'area et un pour le port ce qui dans la pratique limite le nombre de commutateur d'un « Fabric » à 239.

Un « Fabric » correspond à un ou plusieurs commutateurs connectés par un ou plusieurs F_PORTs. Les équipements sont connectés par des N_PORTs.

Le routage se fait selon le Fabric Shortest Path First (FSPF) protocole, les tables de routages sont dans les commutateurs (les tables contiennent d'autres routes que la route optimale pour prendre en compte les pannes possibles). Les tables de routage sont recalculées chaque fois qu'un équipement est ajouté ou retiré ; elles sont dupliquées sur chaque commutateur. La vitesse du lien sert de pondération des liens lors du calcul du plus court chemin.

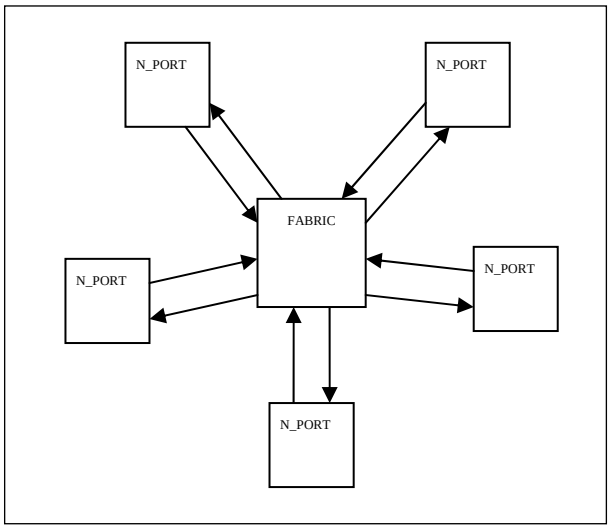

Figure 3-*Fabric*

2.5 Les fonctionnalités des baies de disques Fibre Channel

Une baie FC est composée d'un ensemble de disques qui peuvent être regroupés en grappes, chaque grappe pouvant être définie en RAID. Un Logical Unit Number (LUN) est une partie d'une grappe de disque ; au niveau d'un serveur, un LUN est vu comme un disque. Chaque LUN est affecté au minimum à un contrôleur de la baie.

Un ou plusieurs disques spare peuvent être définis pour pallier à une panne d'un des autres disques de la baie.

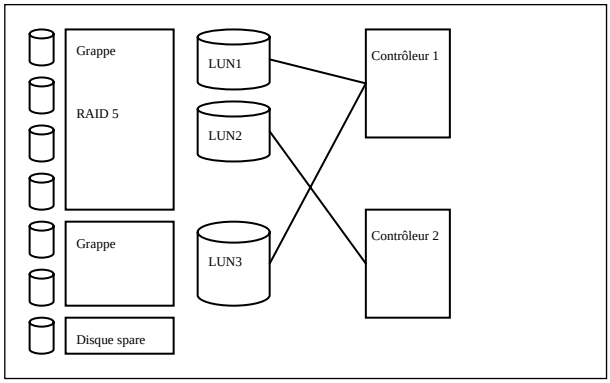

Figure 7 – *Configuration des LUNs*

Trois fonctionnalités avancées fournies par les baies FC actuelles sont intéressantes :

- la copie instantanée
- la réplication synchrone, asynchrone
- le LUN masking

La copie instantanée est un mécanisme intégrée dans le code de la baie, qui travaille en mode bloc, et permet de créer des images d'un LUN en quelques secondes indépendamment de la taille du LUN par recopie de la table des indexes. A la création, elle occupe en moyenne entre 10 à 20% de la taille du LUN initial. Par la suite ne sont recopiés que les blocs modifiés.

La réplication est un mécanisme intégré dans le code de deux baies FC permettant de dupliquer des LUNs d'une baie vers une autre baie identique, située sur un site distant.

Le LUN masking est la possibilité de masquer un LUN pour certains serveurs en fonction de leur WWN.

Problèmes rencontrés :

Si, pour assurer une redondance, un LUN est affecté aux deux contrôleurs, les serveurs voient deux fois le même disque sans le savoir.

Il faut aussi savoir que, de base, le SAN est assez permissif et qu'un serveur peut facilement s'approprier un disque même si celui-ci appartient déjà à un autre serveur. C'est le cas avec Windows qui ne fait pas de contrôle. Ce problème met en valeur tout l'intérêt du LUN masking.

2.6 Interconnexion d'îlots SAN

Une fois un îlot SAN constitué, il est possible de le rendre accessible via le réseau Ethernet.

Les protocoles d'interconnexion sont les suivants :

- Le protocole FCIP

Le protocole FCIP (Fibre Channel over TCP/IP) est une encapsulation des trames FC dans TCP/IP. Tout le trafic (données et gestion) passe par un tunnel ; il y a fusion des deux îlots (l'infrastructure FC est vue et administrée comme un seul réseau de stockage, un îlot).

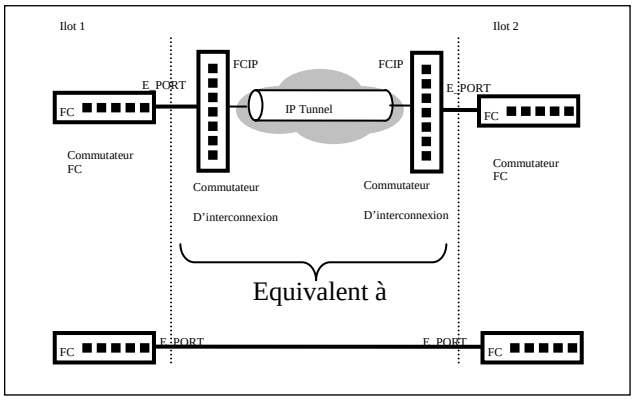

Figure 4- *La connexion FCIP est une extension des E_PORT*

- Le protocole iFCP

Le protocole iFCP (internet Fibre Channel Protocol) permet de relier deux îlots SAN tout en préservant l'indépendance des îlots (gestions indépendantes, erreurs ne se propageant pas d'un îlot à l'autre, sécurité indépendantes) par une topologie multipoint et une translation d'adresses WWN en sessions TCP/IP.

Les ressources rendues accessibles pour l'autre îlot doivent être déclarées au niveau du commutateur d'interconnexion.

- Le protocole iSCSI

Le protocole iSCSI est une encapsulation des trames SCSI dans TCP. Il permet à des serveurs d'accéder à des ressources disque FC via Ethernet : le pilote iSCSI se comporte comme un initiateur pour transporter les requêtes SCSI vers une passerelle IP / FC où se trouve déclarée la cible.

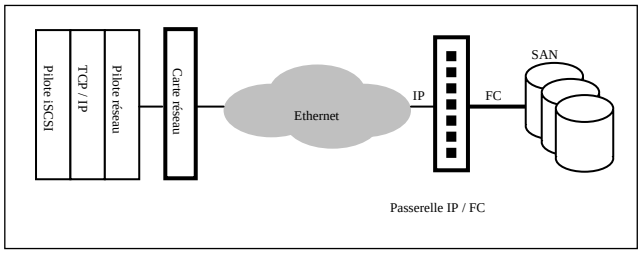

Figure 5- *La connexion iSCSI*

Les ressources sont identifiées par le service de nomage iSNS (internet Storage Name Service).

Des pilotes sont disponibles pour W2000, W2003 et Linux.

3 Le déploiement du SAN expérimental

L'Université Joseph Fourier a réservé dès 2003 des fibres en vue du déploiement d'un réseau de stockage.

Le réseau de stockage contient potentiellement quatre sites distants, plus précisément quatre salles machines du campus universitaire (BIO, ADM-UJF, DSI/GU, UFR/IMA).

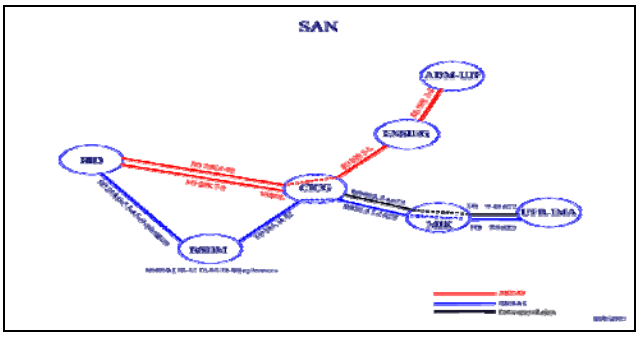

Figure 6 – *Infrastructure fibres réservée pour le SAN multi-site de l'UJF*

Dans le cadre de ce projet, trois baies FC AXUS ELA 1400 ainsi que trois commutateurs FC VIXEL 7100 ont été acquis afin de constituer trois îlots SAN. Tous ces matériels fonctionnent à 1Gb/s.

L'UJF a mis en place un premier îlot pour ses besoins d'espace disque (répertoires personnels des étudiants de l'UFR/IMA).

La DSI/GU a mis en place un deuxième îlot pour expérimentation et la validation des technologies en vue de son utilisation pour des sauvegardes.

Un troisième îlot était en prévision à l'UJF pour le site de BIOLOGIE.

Nos objectifs étaient doubles :

- constituer dans chaque site un îlot avec 1 baie FC, 1 commutateur FC, n serveurs,

- interconnecter les îlots SAN situés dans les bâtiments de l'UFR/IMA, de la BIOLOGIE et de la DSI/GU.

3.1 Le maître mot est : Matrice de compatibilité

Chaque constructeur garantit que son matériel est compatible avec une liste de matériel testé. Et donc décline toute responsabilité en cas de non fonctionnement avec tout autre matériel.

Il faut bien vérifier que toute la chaîne d'accès aux données est compatible. La baie (AXUS ELA 1400) et le commutateur (VIXEL 7100) étaient bien garantis compatibles mais c'est avec les cartes HBA (Host Bus Adapter) FC que nous avons eu des soucis.

1ère mauvaise surprise : nos cartes HBA (JNI) fonctionnaient en AIX 4.3.3 mais n'était plus supportées en AIX 5.1. Et comme nous étions en pleine migration AIX elles sont devenues obsolètes. Heureusement nous avons pu les utiliser pour connecter des serveurs Windows.

2ème mauvaise surprise : alors que nous disposions d'une carte HBA IBM qui marchait, nous avons racheté des nouvelles cartes IBM pour équiper de nouveaux serveurs. Mais la liaison ne se faisait plus entre la carte et le commutateur. Il s'est avéré que les nouvelles cartes à 2Gb/s n'étaient pas capables de s'adapter au 1 Gb/s de notre commutateur. Nous n'avons pas pu obtenir le soutien d'IBM car notre commutateur et notre baie n'étaient pas dans leur certification et nous avons du faire des tests en inter changeant les cartes à chaud pour ne pas perturber la production. Nous avons donc été contraints de nous fournir sur un marché de l'occasion pour trouver des cartes HBA IBM plus anciennes.

3.2 Les différents fonctionnements testés

- Mise en production à la DSI/GU :

Nous avons utilisé dans un premier temps la baie comme suit :

1. Pour réduire la durée des sauvegardes sur bandes LTO2, on passe dans un premier temps par un volume tampon pris sur la baie SAN : les données sont transmises au serveur de sauvegarde (via Ethernet) sur un LUN de la baie. Ceci n'est pas plus rapide que d'écrire directement sur bandes mais le nombre de machines qui peuvent être sauvegardées en même temps n'est pas limité alors que sinon, on est limité à une machine par lecteur. Ceci nous permet d'effectuer toutes nos sauvegardes pendant la nuit. Les données sont ensuite recopiées du tampon vers les bandes LTO2.

Notre espace tampon est de 100Go, les débits sont d'environ 40 Mb/s sur bande comme sur disque.

2. Notre baie sert aussi d'espace de travail pour un de nos serveurs de production qui a une carte HBA compatible avec le commutateur VIXEL (validation de nouvelles versions d'APOGEE, 20Go à chaque fois).

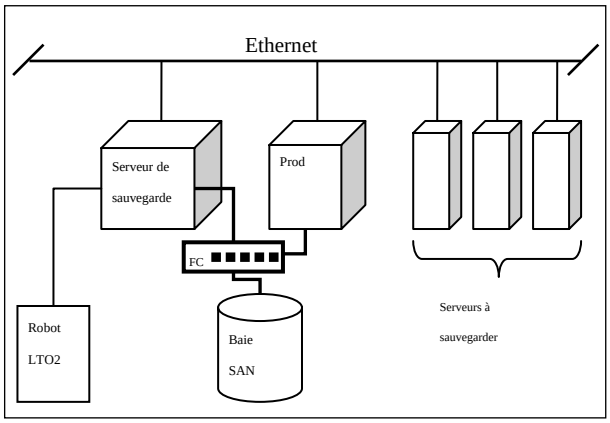

Figure 8 – *Architecture DSI/GU*

Un des futurs objectifs sera de faire transiter les sauvegardes uniquement sur le SAN et plus par Ethernet.

- Mise en production à l'UFR/IMA

Le besoin de l'UFR/IMA est d'offrir à ses étudiants un « home directory » accessible depuis les autres serveurs pédagogiques.

Les données des « home directory » ont été mises sur la baie SAN puis rattachées à un serveur principal et partagées en NFS ou SAMBA pour les autres serveurs.

Cependant même si les serveurs sont tous équipés de carte HBA et reliés au SAN, les données transitent encore par le réseau IP.

En effet, lorsqu'un serveur A possède des données sur une baie et veut les partager avec un autre serveur B, il utilise une couche logicielle (nfs, samba…) pour gérer ce partage. Les données transitent alors par le réseau Ethernet même si le serveur B peut accéder directement à la baie par le réseau de stockage.

Il faut donc dire au serveur B d'aller directement chercher la donnée sur le SAN dans la baie.

Le logiciel SANErgy de chez Tivoli répond à ce problème. Un agent sur chaque serveur intercepte le protocole NFS et redirige les requêtes directement sur lien FC.

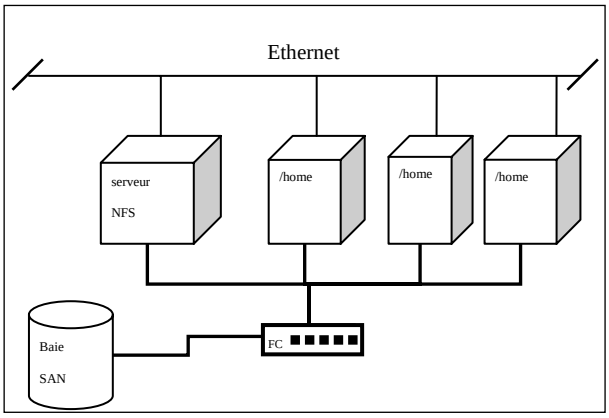

Figure 9 – *Architecture UFR/IMA*

Nous avons fait des tests de performance, les données se transfèrent deux fois plus vite avec SANErgy que sans.

Malheureusement, ce logiciel s'avère lourd car il est dépendant d'un serveur NFS et nécessite l'installation d'un agent sur chaque machine. De plus cette installation nécessite un redémarrage, ce qui n'est pas toujours possible. De plus l'agent est assez instable. Autre lourdeur, SANERgy est multi OS, ses évolutions impliquent donc un redéploiement sur toutes les machines en même temps, de plus elles ne fonctionnent jamais sur les dernières évolutions des systèmes d'exploitation.

Nous avons donc abandonné ce fonctionnement et l'UFR/IMA a utilisé sa baie, via NFS sans l'accélération de SANErgy (c'est-à-dire en fonctionnement de type Network Area Network, NAS).

3.3 L'interconnexion des bâtiments

Cet objectif peut sembler trivial si on fait l'analogie avec le réseau TCP/IP mais les problèmes rencontrés nous montreront que l'interopérabilité du SAN est encore très loin de l'interopérabilité du réseau Ethernet.

L'objectif initial était de relier trois sites, mais pour des questions de facilité de configuration des commutateurs et des baies nous nous sommes limités à deux sites pour réaliser l'interconnexion.

3.4 Problèmes rencontrés

3.4.1 Distance et fiabilité

Même si une des fonctionnalités du SAN mise en avant est de déporter le stockage sur des grandes distances, il n'en reste pas moins qu'il est lié aux mêmes contraintes qu'Ethernet.

C'est-à-dire une fibre multi mode permet une distance de 250m environ, en monomode c'est 10km environ. Ceci dépend aussi du connecteur fibre (Gbic) utilisé.

Il faut préciser que cette distance diminue si le débit du matériel SAN augmente. C'est pourquoi les constructeurs annoncent des très forts débits possibles, mais ne les commercialisent pas encore.

Les sites UFR/IMA et DSI/GU sont distants de 200m environ. Avec des interfaces à 1Gb/s nous sommes proches de la distance maximale ; de plus, le lien comporte beaucoup de jarretières, ce qui a priori entraîne des problèmes de fiabilité dans la connexion.

3.4.2 Problèmes de WWN

Dès lors que la connexion physique a été établie, notre environnement de production s'est cassé. Nous avons donc séparé le SAN de production et le SAN de test d'interconnexion en utilisant le matériel destiné à l'origine pour le troisième îlot.

Il s'est avéré que lorsque nous avions deux baies sur le SAN, celui-ci avait un comportement instable.

Après recherche, nous avons remarqué que les baies de disques avaient toutes les trois le même WWN, alors que nous croyions que cela était impossible : le WWN est censé être unique comme la MAC adresse.

Il est impossible de changer ce WWN directement. Le seul moyen que nous avons trouvé est de changer l'ID des contrôleurs de la baie, ce qui a pour effet de changer le WWN.

En conclusion, nous avons réussi à fusionner nos deux îlots moyennant une administration manuelle des adresses dans l' « Abritrated Loop » que constituent nos commutateurs VIXEL interconnectés.

3.4.3 Problème de « Loop »

Une de nos cartes FC a eu un comportement peu fiable sur le SAN. On s'est aperçu grâce à un pilote de carte FC plus loquace que les autres (sous Linux) qu'en mode « Arbitrated Loop », le commutateur passait son temps à recréer la boucle selon que la carte marchait ou non.

Les équipements FC étant enfichables à chaud, tout ajout ou retrait de matériel provoque des recalculs d'adresses et ceci quelle que soit la topologie.

Il est donc important d'avoir du matériel et des connexions fiables si l'on ne veut pas voir les performances s'écrouler.

3.5 Tests effectués

Sur un serveur Windows2000, nous avons monté des LUNs appartenant à une baie locale DSI/GU et à une baie distante UFR/IMA.

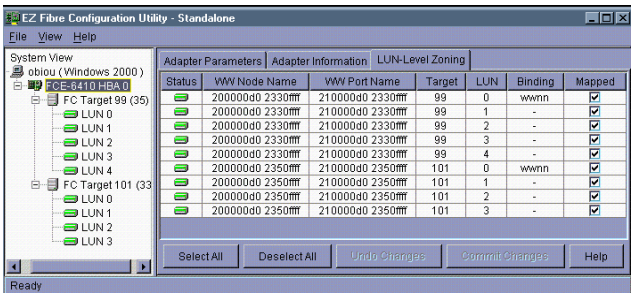

Figure 10 – *Vue du SAN depuis le serveur W2K*

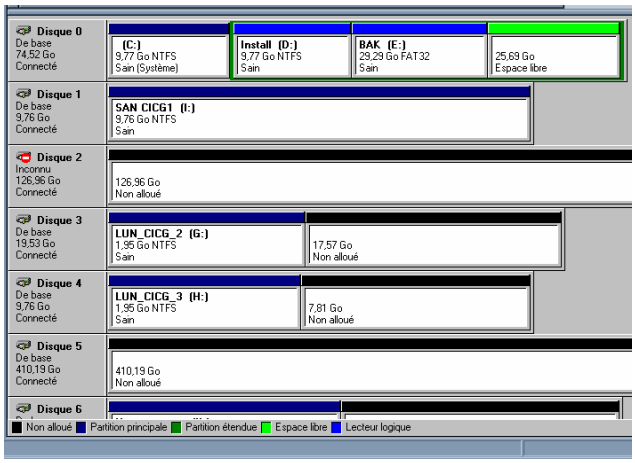

Figure 11 – *Gestionnaire des disques du serveur W2K*

Nous avons transféré de gros fichiers de 600 Mo environ et nous avons mesuré les débits suivants :

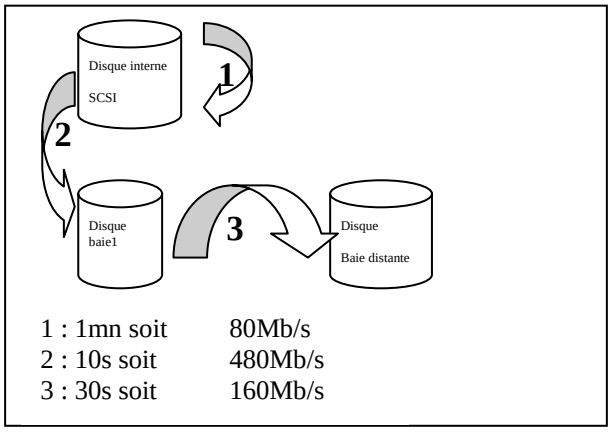

Figure 12 – *Débits mesurés*

3.6 Conclusion pour cette phase de déploiement expérimental

Nous avons rencontré des problèmes techniques semblables à ceux du réseau Ethernet à ses débuts. Depuis les équipements ont bien évolués et présentent maintenant un meilleur niveau d'interopérabilité et d'administration.

Une fois le SAN stabilisé nous avons pu distribuer de l'espace disque aux serveurs connectés au SAN (serveurs AIX, Linux, Windows et Solaris) avec une bonne flexibilité.

4 Solutions opérationnelles

Cette expérimentation du déploiement d'un SAN entre les sites UFR/IMA et DSI/GU nous permet aujourd'hui de pérenniser ces technologies pour solutionner nos problèmes dans le domaine de la sauvegarde et du PRA.

4.1 Site UFR/IMA

La composante UFR/IMA de l'université Joseph Fourier gère l'environnement informatique pour les études de ses 1200 étudiants en informatique, ce qui représente 500 Go de données utilisées de manière intensive : le bâtiment de l'UFR est ouvert de 7h30 à 22h, de plus les étudiants ont accès à leur données à tout moment depuis chez eux. La sauvegarde ne peut plus être effectuée de manière classique sur les serveurs de l'UFR du fait de la durée des fenêtres de sauvegarde (plus de 8h du à une multiplicité de petits fichiers).

L'objectif à atteindre est héberger les données critiques de l'UFR en utilisant la technologie des réseaux de stockage.

La sauvegarde des répertoires des étudiants de l'UFR/IMA de l'UJF et la disponibilité est assurée par réplication sur des baies mutualisées au DSI/GU dans le SAN multi site.

Dans un premier temps, la fonctionnalité de copie instantanée (Snap Shot de l'UFS de Solaris) est utilisée pour avoir un instantané des données des étudiants. C'est la copie instantanée qui servira à la duplication sur la baie distante où à la sauvegarde sur un robot L25 de SUN (25 cartouches LTO2). La sauvegarde est effectuée par le logiciel Networker.

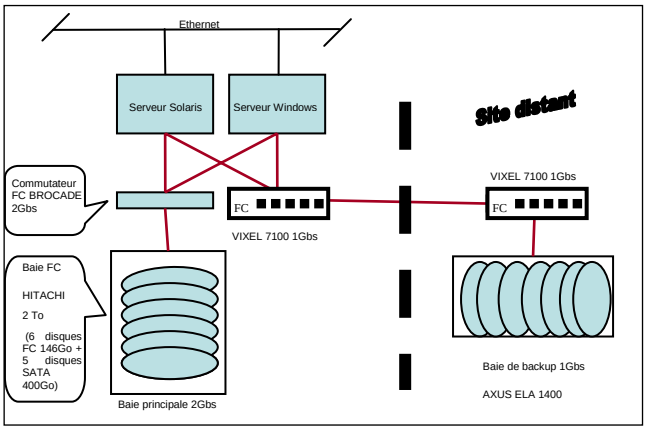

Figure 13 – *Configuration du site UFR/IMA*

Ce projet a été favorablement accueilli par le ministère lors de son appel à projets en octobre 2004.

4.2 Site DSI/GU

Au DSI/GU, le stockage des données est réalisé sur des baies de disques en série attachées aux serveurs (environ 1 Téraoctet de données). Leur sauvegarde et leur archivage sont effectués par le réseau sur un robot de sauvegarde de 60 cartouches LTO2 localisé au DSI/GU et en double par sauvegarde locale (DLT7000 et DLT8000) sur les serveurs. Cette méthode de sauvegarde nécessite une fenêtre de sauvegarde entre 3 et 5 heures. Ce temps d'indisponibilité est maintenant jugé trop long par les établissements pour les applications WEB en amont de la gestion des étudiants et des enseignements.

Le plan de reprise d'activité (PRA) repose actuellement sur une extraction de jeux de sauvegardes hebdomadaires transportées par opérateur sur un site distant (CRI de l'INPG) et transmission toutes les nuits, via le réseau, des logs des transactions des bases données sur ce même site sur un serveur Linux. Toute la documentation qualité concernant la production informatique, en particulier les méthodes de sauvegarde et restauration des données, y est aussi entreposée. Ce plan assure qu'en cas de sinistre majeur du bâtiment du DSI/GU il y ait au maximum une perte de données de 24h, mais il n'assure pas de délai de remise en fonctionnement des applications qui constituent la colonne vertébrale du SI des établissements.

Les objectifs à atteindre en se basant sur les fonctionnalités des réseaux du stockage sont :

- la diminution de la fenêtre de sauvegardes à 5 minutes par la mise en œuvre de copies instantanées sur des baies FC.

- la mise en place d'une réplication des données sur un site distant (UJF domaine universitaire par fibres dédiées et par la suite INPG site Félix Viallet par réseau Ethernet) pour automatiser le PRA du DSI/GU.

Ce projet a également été favorablement accueilli par le ministère lors de son appel à projets en octobre 2004.

A la DSI de Grenoble Université le SI des établissements est implanté sur deux clusters en haute disponibilité. Celui qui héberge l'application de gestion des étudiants et des enseignements Apogée et les applications voisines a été acquis en 2000. Son remplacement a été prévu en 2005. Un marché à procédure adaptée a été lancé pour l'acquisition d'une grappe de serveurs en haute disponibilité avec une infrastructure de stockage basée sur les technologies de réseaux de stockage offrant les fonctionnalités de copies instantanées (snap copies) et de la réplication de baie à baie.

La configuration du nouveau cluster est la suivante :

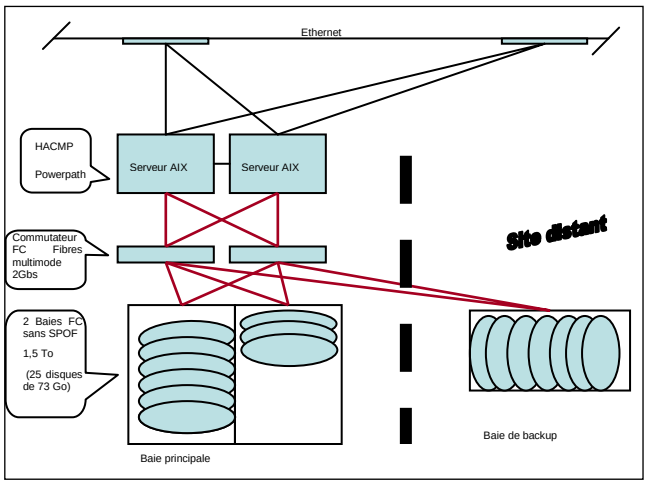

Figure 24 – *Configuration de la grappe de serveurs*

Le matériel a été livré début septembre 2005 et installé dans les deux salles machines sécurisées incendie de la DSI/GU. Le cluster actuel et le nouveau coexisteront pendant 2 mois pour valider les procédures d'exploitation dans le domaine de la haute disponibilité et dans le domaine de la sauvegarde.

Les procédures d'exploitation évolueront en septembre et octobre comme suit :

En ce qui concerne la diminution de la fenêtre de sauvegarde :

- Des copies instantanées des LUNs de la baie FC principale seront déclenchées après shutdown des bases Oracle ; c'est une opération qui ne dure que quelques secondes. Les bases Oracle seront redémarrées après cette opération.

- Ces copies instantanées seront rendues accessibles au système d'exploitation. Leur sauvegarde sera ensuite réalisée par le réseau sur un robot StorageTek (via le logiciel Netbackup) et en local sur dérouleur LTO2 (script basé sur la commande tar), la durée de cette sauvegarde est d'environ 4h. Elle se déroulera pendant que les bases Oracle sont disponibles.

En ce qui concerne le plan de reprise d'activité :

- La réplication des données sur la deuxième baie sera réalisée dans une première phase via la fonctionnalité de mirroring du système d'exploitation (fonctionnalités du JFS d'AIX). C'est la copie instantannée des données qui sera mise en miroir sur la deuxième baie et affectée au système d'exploitation. Elle contient une image cohérente des bases de données Oracle, car faite bases arrêtées. Les archive_log seront répliqués en synchrone sur la deuxième baie par miroir AIX des systèmes de fichiers concernés.

- En cas de sinistre majeur sur le cluster et la baie principale, la baie de backup pourra être connectée en point à point à un serveur sous AIX (après location ou achat) dont le système d'exploitation sera restauré à partir d'un mksysb de nos serveurs de production (symétrique du fait du fonctionnement en HACMP). Les LUNs de la baie FC pourront être réaffectés au système et les points de montages des systèmes de fichiers renommés pour reconstituer l'environnement de production.

Ceci permettra de faire repartir la production en environnement matériel dégradé, mais sans pertes de données dans un délai de temps mesurable (temps de location d'un serveur sous AIX plus le temps de restauration du système plus le temps de la connexion de la baie FC et la réaffectation des LUNs et systèmes de fichier plus le temps de démarrage des bases Oracle).

Nous exposerons dans la présentation, lors des journées JRES en décembre, la mise en œuvre de cette nouvelle procédure de sauvegarde et les éventuels problèmes rencontrés et nous indiquerons jusqu'où le plan de reprise d'activité a pu être réalisé.

5 Conclusions et perspectives

Dans le domaine des technologies de stockage les investissements et infrastructures sont lourds. De plus, toute modification ou interconnexion est à gérer avec précaution car pouvant créer des disfonctionnements de l'existant. Il est donc impératif de définir l'usage avec précision avant d'envisager un déploiement.

Néanmoins, les fonctionnalités apportées par une infrastructure de réseau de stockage sont très intéressantes.

Une fois mis en place, l'infrastructure SAN fait preuve d'une remarquable stabilité tant au niveau réseau qu'au niveau stockage, ce qui pose cette technologie en candidat incontournable pour toutes les applications critiques.

En second lieu, les performances sont au rendez vous, même si elles n'atteignent pas les sommets annoncés (les temps d'accès à des fichiers sont divisés par deux entre un accès par Ethernet - NFS ou CIFS - et un accès direct par carte FC à 1Gb/s).

En troisième lieu, une fois mis en place, le SAN permet de mettre à disposition très rapidement des espaces disques sécurisés de taille choisie, pour n'importe quel OS. Nous avons pu ainsi nous en servir au pied levé pour des VSR d'Apogée et pour accélérer le fonctionnement de nos sauvegardes via les réseaux avec le logiciel Netbackup.

Pour l'année à venir, nous travaillons à transformer notre SAN multi site expérimental en un réseau de stockage « campus wide » qui respecte l'autonomie des établissements de l'académie de Grenoble. Ce SAN nous permettra de proposer aux établissements grenoblois des services liés au stockage en assurant une forte disponibilité et une grande qualité de service (ces services seront disponibles en FC et en iSCSI). Nous visons notamment des services de sauvegarde et d'archivage pour la mise en place ou l'amélioration des plans de continuité de service des applications vitales pour les universités.

Bibliographie

[1] John Vacca, *The essential Guide to Storage Area Networks* Prentice Hall PTR 2002.

[2] McData, Protocole IP et Routage FC, Journée de formation, Paris, Mars 2004.