Solution de stockage répartie sur les centres de recherche INRIA à base de serveurs de fichiers de type « NAS »

Laurent Mirtain
INRIA Sophia Antipolis - Méditerranée
2004 Route des Lucioles – BP93 – 06902 Sophia Antipolis Cedex
Laurent.Mirtain@sophia.inria.fr

Jean-Luc Szpyrka
INRIA Sophia Antipolis - Méditerranée
2004 Route des Lucioles – BP93 – 06902 Sophia Antipolis Cedex
Jean-Luc.Szpyrka@sophia.inria.fr

Résumé

L'INRIA (Institut National de Recherche en Informatique et Automatique) s'est doté en 2006 d'une solution de stockage répartie à base de serveurs de fichiers de type NAS[1]. L'architecture de cette solution est conçue autour d'un équipement central vers lequel sont répliqués les fichiers des serveurs NAS déployés dans ses centres de recherche. Ce serveur de réplication constitue la base d'un plan de reprise d'activité (PRA), qui doit permettre en cas d'indisponibilité d'un équipement de site de récupérer, voire d'utiliser à distance les données répliquées sur le NAS central.

A travers cet article, nous décrirons les configurations mises en place, leur intégration dans notre infrastructure, le PRA, les modes d'utilisation de ces équipements (partage de fichiers, accès en mode bloc FCP[2] ou iSCSI[3], hébergement de données de services applicatifs, dont des bases de données). Nous ferons le bilan d'un an et demi d'exploitation de la solution, notamment sur le niveau de disponibilité des équipements et l'efficacité du PRA. Enfin nous évoquerons les évolutions envisagées pour améliorer notre solution.

Mots clefs

NAS, NFS, NFSv4, CIFS, iSCSI, FCP, LDAP, Kerberos, réplication, PRA

1 Genèse du projet

Courant 2004, la direction informatique de l'INRIA établit un état des lieux des solutions de stockage et de sauvegarde déployées dans ses différents centres. Il en ressort que chaque site utilise des solutions très dissemblables, allant d'une centralisation totale des fichiers utilisateurs, à l'éclatement des données sur plus de 900 postes de travail, tous sauvegardés. Les équipements existants (serveurs de stockage, robots de sauvegarde...) arrivant en fin de vie, il était également opportun de définir une solution de remplacement commune permettant :

- de centraliser les données importantes ;
- de diminuer le nombre de serveurs ;
- de simplifier la sauvegarde et d'améliorer la vitesse de restauration des données critiques ;
- d'offrir un meilleur niveau de sécurisation et de consolidation ;
- d'établir formellement un plan de reprise d'activité.

Fort de l'expérience du centre de Nancy, qui utilise depuis fin 2003, deux serveurs de fichiers NAS, un groupe de travail, constitué début 2005, est chargé d'étudier l'opportunité d'une acquisition groupée de ce type de matériel. Cette étude s'appuie sur l'état des lieux de 2004, qui précise l'environnement technique des différents centres INRIA :

- caractéristiques des données : volumétrie, criticité, type de sécurité (Unix ou Windows) ;
- nombre de clients simultanés potentiels ;
- modes d'accès NFS et CIFS depuis les OS Linux, Windows, MacOS ;
- caractéristiques des équipements LAN et des interconnexions WAN à travers le réseau RENATER[4] ;
- bases d'authentification utilisateurs existantes (LDAP, NIS et Active Directory) ;
- type d'équipement de sauvegarde.

Le groupe a effectué tout d'abord, une étude du marché sous la forme d'une veille technologique, de rencontres avec les principaux acteurs et d'échanges avec d'autres établissements. Il s'est intéressé aux fonctionnalités avancées des équipements, notamment l'accès en mode bloc, la sauvegarde sur disques des fichiers et leur réplication, les fonctionnalités d'administration et l'interopérabilité avec nos services en place (LDAP, NIS, Active Directory...). Il a réalisé une maquette, grâce au prêt d'un NAS supplémentaire, qui a permis de tester une réplication inter sites et l'accès distant aux fichiers répliqués depuis un client sous Linux et Windows.

En conclusion de son étude, le groupe a préconisé d'équiper les centres de serveurs NAS et de lesconsolider

[1] NAS : Network Attached Storage

[2] FCP : Fibre Channel Protocol

[3] ISCSI : Internet SCSI, protocole de transport de SCSI sur TCP/ IP.

[4] RENATER : Réseau National de télécommunications pour la Technologie et l'Enseignement et la Recherche

par une solution de réplication constituant le PRA, avec trois possibilités de scénarios techniques :

1. réplication circulaire : le site A réplique le site B, qui réplique le site C...qui réplique le site A ;
2. réplication en étoile sur un site central ;
3. réplication locale sur chaque site (sans réplication centrale).

Le scénario 2 a été retenu car c'est le plus économique au delà de 2 sites. Il est évolutif (facilité d'intégration d'un nouveau site), il permet de n'utiliser qu'une seule solution de sauvegarde sur bandes localisée sur le site central et il offre une solution de PRA en cas de sinistre grave sur un site ; il présente toutefois l'inconvénient d'exiger une bonne inter connectivité réseau entre sites et de devoir coordonner l'exploitation de la solution globale.

Le groupe a été ensuite chargé de la mise en place du marché d'acquisition. Compte-tenu du coût potentiel des équipements attendus, le marché initial prévoyait des tranches fermes et conditionnelles et des options d'achats, permettant ainsi, selon le tarif, l'acquisition d'équipements supplémentaires. Le marché a été notifié en décembre 2005, la livraison, l'installation des matériels, la mise en service de la réplication se sont déroulées en janvier et février 2006. Un marché complémentaire, mis en place fin 2006, a permis d'accroître la volumétrie des NAS.

2 Notre solution NAS

2.1 Description de la configuration

La figure 1 représente l'infrastructure des serveurs de fichiers NAS en place à l'issue du marché et le tableau ci-après détaille les configurations matérielles sur chaque site.

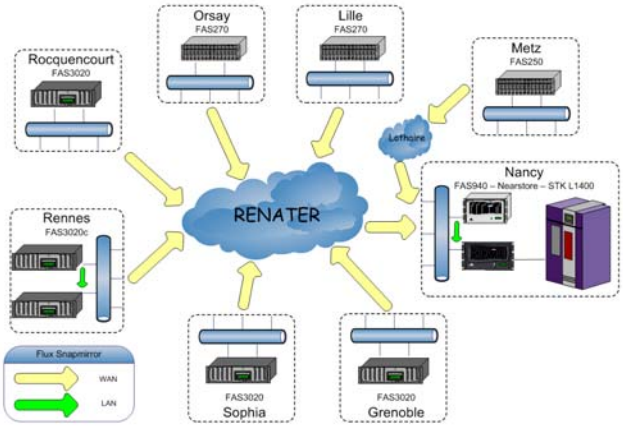

Figure 1 - *La solution de stockage*

Site	Modèle, capacité brute (type de disques)
Grenoble	FAS 3020, 16 TB (FC)
Lille	FAS 270, 6 TB (FC)
Metz	FAS 250, 144 GB (FC)
Nancy	FAS 940, 4 TB (FC) Nearstore R200 56 TB (ATA) Librairie de bandes StorageTek L1400 de 300 Slots avec 3 lecteurs SDLT 600
Orsay	FAS 250, 1TB (FC)
Rennes	FAS 3020 cluster, 18 TB (FC) + 14 TB (SATA)
Rocquencourt	FAS 3020 20 TB (FC)
Sophia	FAS 3020 17 TB (SATA)

Tous les serveurs NAS fournis (appelés également *filers*) sont des matériels NetApp. Leurs principales caractéristiques sont :

- technologie RAID matérielle double parité (DP) spécifique, tolérant la panne simultanée de deux disques, couplée à des disques de « hot spare » ;
- possibilité d'avoir sur le même NAS des groupes RAID de disques performants de technologie Fibre Channel et des groupes RAID de disques capacitifs de technologie SATA ;
- redondance des contrôleurs disques, des alimentations électriques, des interfaces de connexion réseau, des bus d'accès aux disques ;
- changement et ajout de composants à chaud ;
- extensions en capacité de stockage par ajout de modules disques, en les intégrant ou non aux volumes existants ;
- système d'exploitation spécialisé (Data ONTAP®) administrable via des commandes en ligne ou une interface Web ;
- support natif des protocoles NFS, CIFS , NIS , Active Directory , NTP , LDAP, iSCSI , Fiber Channel, IPSec ;
- système de fichiers « Write Anywhere File Layout » (WAFL®) avec possibilité de redimensionner les volumes à chaud, mécanisme de quotas, support des ACLs CIFS ou NFS ou mode mixte ;
- mécanisme intégré de sauvegarde sur disques permettant la restauration par l'utilisateur ;
- possibilité de réplication des volumes sur un autre filer ;
- mécanisme de reboot rapide (moins de 2 minutes).

Les NAS s'appuient sur les serveurs d'authentification locaux à chaque centre : Active Directory pour l'accès aux volumes CIFS, NIS pour l'accès en NFS. Nous verrons plus loin dans l'article, la solution LDAP+Kerberos, utilisée par un des centres INRIA.

Il faut noter, qu'entre les groupes RAID, les disques de SPARE, le formatage WAFL, le découpage en LUNs et les réservations pour les sauvegardes (prévoir 20% du volume du LUN), la taille disque réelle disponible ne représente plus que la moitié de la taille brute initiale (une capacité de 18 TB brute correspond à une taille utile de 9 TB).

2.2 Sauvegarde et réplication

Un des points forts de ces serveurs NAS est leur solution de sauvegarde (logiciel SnapShots®). Elle utilise un mécanisme de copie instantanée appelé « snapshot. »

Créer un snapshot consiste à dupliquer uniquement les blocs d'inodes, alors que les blocs de données qui sont identiques entre le volume actif et le snapshot, eux, ne sont pas dupliqués. C'est seulement avant d'être modifié ou supprimé du volume actif qu'un bloc de données est affecté au snapshot et qu'un nouveau bloc de données est associé au volume actif (Figure 2).

Le temps nécessaire pour fabriquer le snapshot d'un volume est de l'ordre d'une minute, car il ne dépend que du

nombre de répertoires et de fichiers et pas de son volume ; il est, de plus, peu consommateur d'espace disque (son volume croît en fonction du nombre de blocs de données modifiés) et perturbe peu l'activité I/O en cours (copie des blocs d'inodes uniquement).

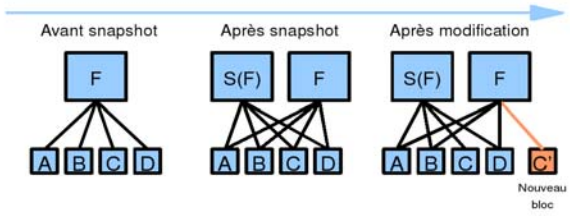

Figure 2 - *principe des snapshots*

Ces caractéristiques nous permettent de produire huit snapshots horaires sur la journée en cours, d'en conserver les dix derniers ainsi que les trente derniers snapshots quotidiens, déclenchés, eux, à minuit. Nous conservons ainsi, pour nos volumes de fichiers utilisateurs, une quarantaine de snapshots qui engendrent un surcroît d'environ 20% du volume total des données (valeur significative d'un environnement de type « homedir » dans lequel les fichiers évoluent peu).

Les utilisateurs peuvent accéder directement en NFS ou CIFS au contenu des snapshots via un répertoire *.snapshot* et restaurer eux-mêmes leurs fichiers.

La réplication se fait, elle, via le logiciel SnapMirror®. Il permet de réaliser la copie de volumes de filer à filer au travers d'un réseau TCP/IP, LAN ou WAN. Ce logiciel doit être installé sur les filers source et cible. Ce dernier, doit disposer de la volumétrie suffisante pour recevoir la totalité des copies. Dans le cas du WAN, qui est le nôtre, la réplication est asynchrone.

Le principe de fonctionnement de SnapMirror est le suivant :

- déclenchement d'un snapshot pour construire la liste des blocs à répliquer : à l'initialisation de la réplication, la copie est complète, puis les fois suivantes, elle se fait de manière incrémentale, bloc par bloc ;
- lancement et contrôle du flux de réplication vers le NAS miroir avec reprise de la synchronisation, en cas d'échec.

2.3 Équipe d'exploitation

L'équipe en charge de l'exploitation comprend un ou deux correspondants par centre, soit une douzaine de personnes, dont un animateur. Elle est organisée autour d'une liste de diffusion et d'une réunion de suivi (par audio ou visio conférence) mensuelle. Elle assure l'administration de la solution et sa maintenance, ainsi que les relations de suivi avec le prestataire du marché.

Un mécanisme de remontée d'information par email (*AutoSupport*) permet à la maintenance du constructeur de recevoir tous les évènements du système et de produire un bilan de santé des serveurs (consultable sur un site Web) qui facilite le suivi des serveurs.

3 Le PRA

3.1 La théorie...

Notre PRA doit nous permettre :

- de reconstruire les volumes d'un filer de site en cas de panne matérielle, logicielle ou d'erreur humaine ;
- de restituer les données à partir des sauvegardes sur bandes en cas d'une corruption de données sur le filer de site et sur le NAS de réplication ;
- d'offrir un mode d'accès distant sur le NAS réplica en cas de panne prolongée du filer de site.

Le passage en mode d'accès distant, puis le retour en mode normal, sont des opérations complexes (voir §3.1.3). Nous avons donc convenu du principe de mise en œuvre suivant :

- si l'indisponibilité des données d'un site est estimée inférieure à la journée et si ces données ne sont pas perdues (panne matérielle par exemple), le PRA est activé en donnant l'accès en lecture seule aux données répliquées, le temps de remettre en service le NAS du site ;
- sinon, le PRA est activé en donnant l'accès en lecture/écriture aux données répliquées. Après réparation, les données devront être synchronisées en sens inverse vers le filer source avant le retour à la normale.

3.1.1 Configuration de la réplication

Lorsqu'elle est asynchrone, il faut réduire l'intervalle entre deux réplications, pour minimiser la perte potentielle de données. Pour déterminer cet intervalle, il faut prendre en compte le taux de modification des données, la bande passante du réseau, la taille du volume à répliquer et le nombre de flux de réplication.

Notre PRA actuel comporte 37 volumes répliqués, dont la taille cumulée par réplication est d'environ 10 GB. Nous avons choisi une fréquence de réplication de 1 heure, ce qui représente un taux de transfert continu d'environ 20 Mbits/sec. Sachant que le volume global répliqué est de l'ordre de 13 TB, le taux du volume transféré est d'environ 0,07% par heure et 1,8% par jour.

3.1.2 Configuration de l'accès distant

Le filer central sert de secours aux filers de sites qui ont chacun leurs propres bases d'authentification. Comme un filer ne peut pas utiliser simultanément plusieurs bases d'authentification pour un protocole donné (CIFS ou NFS), NetApp fournit en option, le logiciel de virtualisation MultiStore®. Il permet de « découper » un NAS physique en plusieurs NAS logiques. A chaque filer de site est ainsi associé sur notre file central, un filer « virtuel » (appelé vfiler) qui dispose de ses propres volumes disques, interfaces réseau et domaines de sécurité.

Pour permettre, le cas échéant, aux postes clients distant d'un site, de se connecter au vfiler de secours, des défiltrages appropriés sur nos routeurs de sites permettent à chaque vfiler d'accéder aux services d'authentification NFS/CIFS du site auquel il sert de secours. Pour faciliter ce

basculement des clients, nous utilisons un mécanisme d'alias DNS qui permet d'associer le nom du filer soit au filer de site, soit au vfiler. En cas de basculement sur le NAS de secours, il suffit de changer l'enregistrement DNS de cet alias pour qu'il pointe sur l'adresse du vfiler correspondant. Le basculement de filer étant une procédure manuelle relativement longue (voir ci après), ce mécanisme d'alias suffit.

3.1.3 Procédures de PRA

Le déclenchement du PRA se fait via des procédures manuelles établies par notre prestataire. Les principales étapes sont :

- arrêter la réplication ;
- activer le vfiler et changer l'enregistrement de l'alias DNS pour faire pointer les clients vers l'adresse du vfiler ;
- accompagner le basculement des postes clients (dans la pratique le reboot est quasiment nécessaire pour purger le cache DNS du client) ;
- réparer le NAS défaillant, puis ;
- établir un miroir dans l'autre sens pour répliquer les données modifiées, ce qui peut prendre un certain temps si le filer doit être complètement reconstruit...
- programmer une interruption sur le site pour lancer une dernière synchronisation puis arrêter la réplication inverse ;
- re basculer sur le filer de site, une fois opérationnel, après modification de l'alias DNS ;
- accompagner le basculement des postes clients (reboot)...

En régime normal, l'équipe d'exploitation doit maintenir le PRA opérationnel en vérifiant que tous les volumes à répliquer sont bien déclarés et que les réplications se déroulent correctement.

3.2 ... et la pratique

Au delà du discours théorique, il y a la réalité du terrain... à savoir celle de nos liaisons WAN RENATER.

La **réplication** ne pose pas de problème et se montre parfaitement fiable. A ce jour, nous avons 37 flux de réplication répartis sur une heure (fréquence de réplication). La figure 3 donne un aperçu du trafic d'une partie des flux de réplication, transitant sur une des trois interfaces Gigabit Ethernet du filer central.

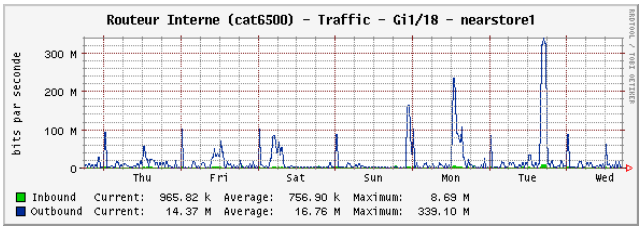

Figure 3 - flux de réplication.

Il n'en est pas de même avec la **reconstruction**. Lors du maquettage, puis de l'installation des NAS, nous avons testé les performances de nos liaisons inter sites avec le logiciel iperf et les avons comparés avec celles obtenues par la réplication et le travail à distance. Les résultats de iperf[5] sont compris entre 8 et 28 ms pour la latence et environ 200 Mbits/sec pour la bande passante (conforme à notre agrément RENATER). Les débits applicatifs obtenus par la réplication, ont donné des résultats variant de 50 à 100 Mbits/sec selon le site, soit dans le meilleur cas, 24 heures par TB.

Compte-tenu des volumétries de nos filers, il n'est pas possible d'utiliser le réseau WAN pour les reconstruire dans un délai raisonnable. Pour réduire ce délai, les débits de réplication pourraient être améliorés en optimisant les paramètres systèmes des filers (fenêtre TCP) ; mais ces flux devant partager la bande passante avec d'autres trafics applicatifs, nous n'avons pas poussé l'investigation sur ce point : le gain obtenu serait insuffisant et ne résoudrait de toute manière pas la question de la reconstruction via le WAN. Par ailleurs, des politiques de QoS déployées sur certains de nos routeurs de site limitent le débit maximum possible pour un flux donné.

Des solutions alternatives consisteraient à :

- utiliser le logiciel SnapMirror to tape (SM2T) fourni par NetApp qui permet de reconstruire à partir de bandes (il faut disposer alors de l'infrastructure matérielle pour produire, transporter, puis décharger les bandes sur le filer) ;
- utiliser les services du prestataire, qui propose une prestation de transfert via un filer « intermédiaire » transporté par véhicule d'un site à l'autre ;
- disposer d'un NAS de réplication en local (c'est la solution qu'a choisi le centre de Rennes).

Pour ce qui est du **travail à distance**, des tests de copie de fichiers en NFS et en CIFS et de calcul de la taille d'une arborescence (du -ks), effectués avec quatre utilisateurs connectés simultanément sur leur homedir, ont révélé des durées en mode distant entre 10 et 20 fois supérieures au mode local. Ces résultats montrent clairement que les protocoles CIFS et NFS ne sont pas adaptés aux temps de latences élevés et aux métriques de nos liaisons WAN.

Nous avons envisagé les solutions suivantes :

- l'optimisation des paramètres système. Une étude [1], réalisée dans le cadre de Grid'5000, montre qu'il est possible de booster les performances de TCP d'un OS Linux sur des liens WAN. Mais ce qui est envisageable sur des machines d'expérimentation n'est pas transposable sur un parc de plusieurs centaines de postes de travail ;
- l'utilisation de boîtiers WAFS[6]. Ces boîtiers accélèrent le trafic NFS/CIFS entre sites distants, mais cette solution est économiquement peu viable dans le cadre d'une utilisation ponctuelle (le cache étant vide au moment du basculement) ;
- là aussi, disposer d'un NAS de secours en local, permettrait d'avoir une solution de bascule utilisable.

[5] http://fr.wikipedia.org/wiki/Iperf

[6] La technologie WAFS (Wide Area File Services) permet d'optimiser les accès distants vers un serveur de fichiers NFS/CIFS via des mécanismes de lecture anticipée, de cache local et de compression à la volée.

Des pannes inattendues : Les deux gros soucis que nous avons connus, sont venus de problèmes logiciels. Dans un cas, un bogue dans l'OS a provoqué, sur la simple corruption d'un fichier, le plantage brutal d'un filer avec impossibilité de redémarrage, et dans l'autre, un bogue dans la gestion des buffers en écriture NFS a fait tomber les interfaces réseau du filer. Ces deux bogues sont enregistrés officiellement chez NetApp (BugID 245282 et 248621**)** mais n'ont toujours pas de correctif à ce jour. Nous restons donc vulnérables...

Dans un cas, l'interruption a duré 8 heures, mais dans l'autre, le filer étant en configuration cluster (actif-actif), il a suffit de basculer tous les volumes sur le deuxième filer, encore opérationnel.

3.3 Bilan du PRA

Le PRA n'est pas, en l'état actuel, utilisable comme nous l'envisagions initialement : il n'est pas possible de travailler à distance dans des conditions de performances acceptables et les volumétries des filers ont atteint de telles tailles, que leur reconstruction par le réseau serait trop longue.

En l'état actuel, le PRA fournit une solution de reprise sur sinistre grave (incendie, inondation...) et permet d'utiliser une solution de sauvegarde centralisée Pour être pleinement opérationnel nous devons mettre en place une réplication locale, comme c'est le cas à Rennes par exemple.

4 Modes d'utilisation de nos NAS

4.1 Service de fichiers utilisateurs

Nos filers sont avant tout des serveurs de fichiers, avec les points forts que sont le support de CIFS natif, la sauvegarde horaire par snapshots et la possibilité de restauration de fichiers par l'utilisateur.

Ce mode d'utilisation concerne l'essentiel des données hébergées sur nos filers : homedirs utilisateurs, profils itinérants, espaces partagés par les équipes de recherche ou les services de support, données d'archivage ou temporaires. Il représente environ 60% de la volumétrie globale de nos serveurs.

4.2 Service de fichiers applicatifs

Compte-tenu de leur niveau de disponibilité, de performance, de la sauvegarde intégrée et du PRA, nous avons tout intérêt à héberger sur nos NAS, nos données applicatives importantes. Il faut toutefois vérifier que l'application supporte l'utilisation d'un volume NFS, en terme de fonctionnement et de performance. NetApp [2] fournit des recommandations sur l'utilisation de ses filers dans ce cadre. Celles-ci indiquent, selon l'application, comment optimiser les options de montage NFS, les paramètres systèmes du client et du filer, l'organisation des fichiers sur le NAS et la configuration réseau.

Optimiser le comportement du client NFS : cache, mécanismes de lecture anticipée (*prefetch*) ou d'écriture différée (*write-back*), gestion des verrous (*lock*), il faut trouver le bon compromis performance/rafraîchissement pour que le serveur NFS ne diffère pas trop les modifications, sans quoi les autres clients ne les verraient pas. Augmenter la taille des transferts I/O de NFS permet de réduire l'*overhead* réseau du trafic NFS (en limitant la fragmentation des paquets). Régler le nombre simultané et la taille des transferts possibles dans la pile TCP permet aussi d'en booster les performances.

Optimiser la configuration réseau : liens GbE[7], agrégats, auto négociation de bout en bout, Jumbo frames[8], isolement du trafic data sur un VLAN « stockage » dédié sont autant de facteurs favorisant les performances de NFS. Par ailleurs, utiliser NFS en mode TCP et l'option de montage *hard* permettent à l'application de rester en attente si la connectivité au serveur est perdue.

La figure 4 montre notre mode de raccordement au réseau d'un service applicatif avec données sur le NAS

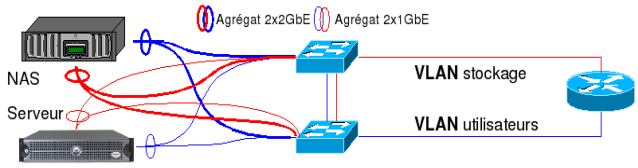

Figure 4 : *raccordement au réseau*.

4.2.1 Hébergement d'un serveur Apache en NFS

Une des premières applications dont nous avons hébergé les données sur le NAS est un serveur Web Apache. La documentation NetApp [2] ne fournit aucune information sur l'utilisation de ses filers pour un serveur HTTP Apache. Par contre, la documentation Apache 2.x [3] donne des préconisations concernant l'utilisation d'un volume NFS (directives à appliquer dans le fichier de configuration httpd.conf). Il y est indiqué :

- d'écrire les fichiers de lock sur le disque local et pas sur le disque NFS (directives **LockFile** et **RewriteLock) ;**
- de désactiver les mécanismes noyau *memory mapping* et *sendfile* (directives EnableMMAP et EnableSendfile).

En suivant ces recommandations, le fonctionnement d'Apache avec données sur le NAS donne toute satisfaction. A noter que nous utilisons la réplication SnapMirror, pour dupliquer le contenu d'un site Apache sur le filer d'un autre centre et construire ainsi un service Web réparti sur plusieurs serveurs.

4.2.2 Hébergement de SGBD en NFS

NetApp certifie, au travers de matrices de certification [4] ou de guides « Best Practices » [5], la compatibilité de ses filers avec les principaux SGBD du marché (Oracle, Sybase, SQL Server, MySQL Enterprise Edition...).

Utilisation pour une Base Oracle 10g : La deuxième application, que nous avons hébergée, est une base Oracle utilisée par l'application Oracle Calendar (agenda partagé). Nous avons appliqué les recommandations de NetApp concernant les options de montage NFS pour Oracle [6], mais, compte-tenu des faibles volumétrie et activité I/O de notre base (l'application Oracle Calendar n'utilisant pas intensivement le backend Oracle 10g), nous n'avons pas séparé les données et les journaux sur différents volumes.

[7] GbE : Gigabit Ethernet

[8] Jumbo frames : trames Ethernet de taille supérieure à 1500 octets.

Notre service Oracle fonctionne sans soucis depuis novembre 2006, repartant de bon pied, même après une coupure intempestive de l'accès à ses fichiers.

Utilisation pour une Base MySQL : Nous avons ensuite hébergé sur un NAS une base MySQL utilisée par des services Web. Cette instance n'est accédée que par un seul serveur à la fois et utilise le moteur, non transactionnel, MyISAM. Contrairement à NetApp, le manuel de référence de MySQL5, [7] dissuade d'utiliser NFS pour stocker les fichiers de données et les journaux, particulièrement avec le moteur InnoDB (MySQL mettant en avant des problèmes potentiels liés aux verrous NFS des fichiers). Nous avons fait de nombreux tests avec les suites *sql-bench* et *crash-me*[9] pour éprouver notre configuration et comparer les performances entre l'utilisation du disque local et du NAS. Les résultats se sont montrés satisfaisants avec le NAS et nous avons mis en service cette configuration. La Figure 5 montre un comparatif du test *sql-bench* entre MySQL utilisant le disque local (SCSI à 15000 tpm) et Mysql utilisant NFS sur le NAS (disques SATA à 7500 tpm) pour une même machine (Dell PE2850, RHELv4, filesystem ext3).

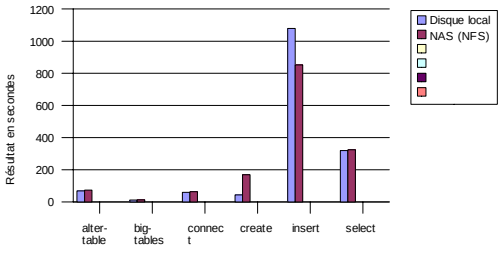

Figure 5 - *comparaison SCSI/NAS(NFS) pour sql-bench*

4.2.3 Sauvegarde en iSCSI

Nos filers permettent une utilisation en mode bloc via iSCSI ou FCP, qui peut être une alternative lorsque le mode NFS n'est pas supporté par l'application. Nous n'utilisons Fibre Channel que pour raccorder la librairie de bandes au NAS central ; sinon, nous privilégions iSCSI car il est plus simple à administrer, ne nécessite pas d'équipements matériels supplémentaires (utilisation des équipements réseau) et les compétences qui vont avec...

Il est utilisé sur 2 sites, dans le cadre de la sauvegarde sur disques, avec deux serveurs, un en Linux RHEL v4 et l'autre en Windows 2003 Server. Sur les recommandations de Dell, nous n'avons pas installé de carte matérielle iSCSI, mais rajouté une carte réseau « dual port Gbit Ethernet » additionnelle, agrégée à 2 Gbps au réseau de stockage (Figure 4). Nous utilisons les initiateurs iSCSI fournis en standard avec les OS.

Pour Linux, NetApp [8] conseille, pour des questions de performances, d'augmenter la taille de la file d'attente du driver iSCSI, côté serveur et côté NAS, et surtout, de partitionner le LUN iSCSI en alignant sa géométrie avec celle des blocs de 4096 octets de WAFL.

Dans les faits, iSCSI donne toute satisfaction avec les configurations que nous utilisons. Depuis le noyau 2.6, le serveur Linux résiste à une coupure du réseau de stockage : les I/O se mettent en attente puis reprennent dès que le disque est à nouveau visible.

Vous trouverez sur le Web des articles très complets, comparant les performances de NFS et de iSCSI sur des plate-formes de tests perfectionnées [9]. Pour notre part, nous avons comparé les performances en accès séquentiel[10] entre iSCSI et NFS avec le logiciel de sauvegarde Networker (Figure 6), et avec la commande dd - en y ajoutant la comparaison avec le disque local (Dell PE2850, RHELv4, filesystem ext3, disques SCSI à 15000 tpm) - (Figure 7). Ces tests ont été effectués alors que notre NAS était en production, les résultats obtenus reflètent des performances en usage « réel. »

Débit en MB/s	NFS	NFS optimisé	iSCSI
Sauvegarde (3 en //)	40	47	58
Restauration	45	57	44

Figure 6 -C*omparatif NAS NFS/iSCSI avec Networker*

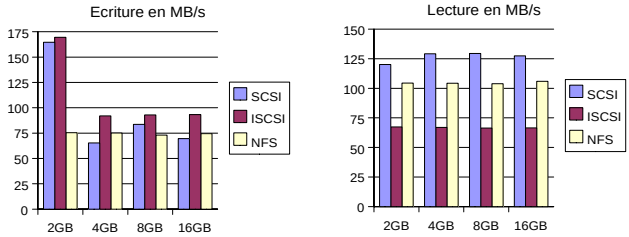

Figure 7 -C*omparatif SCSI/NAS iSCSI/NFS avec dd*

Nos résultats montrent des performances en lecture séquentielle NFS supérieures à celles obtenues en iSCSI. Compte-tenu des conditions de tests, il est difficile de faire une analyse poussée, mais on pourrait en conclure que l'implémentation de NFS sur le NAS (et/ou sur le client Red Hat) est plus optimisée que celle de iSCSI.

Des tests réalisés par NetApp confirment en tout cas, la proximité des résultats obtenus entre NFS et iSCSI [10]. Par ailleurs, même si nous avons appliqué les recommandations de NetApp pour iSCSI, nous n'avons pas cherché à optimiser nos résultats, qui sont suffisants dans le cadre de la sauvegarde sur disques. Nous reviendrons sur ce chantier, certainement lorsque nous voudrons utiliser iSCSI pour une application plus critique (nous projetons par exemple d'héberger sur un filer en iSCSI nos images virtuelles XEN ou VMware [11]).

4.3 NFSv4

NFSv4 [12] a été mis en place sur les NAS du centre de recherche Futurs (Lille et Orsay). L'objectif est d'utiliser les fonctionnalités avancées de NFSv4 comme l'accès sécurisé et le mécanisme de disponibilité des données hors connexion.

NFSv4 a besoin d'un service d'authentification Kerberos[11], couplé à une base utilisateurs (ici, un annuaire LDAP). Le

[9] Sql-bench et crash-me sont des suites de tests fournies avec MySQL

[10] A noter que l'utilisation des Jumbo frames n'a pas donné une amélioration significative.

[11] Kerberos est un système d'authentification sécurisé à base de jetons.

serveur Kerberos (le KDC, Key Distribution Center) déployé est celui de Heimdal[12]).

Comme nos serveurs NAS ne peuvent être associés qu'à un seul domaine Kerberos, et qu'ils utilisent celui de Active Directory (AD) pour l'accès CIFS, il a été mis en place une relation d'approbation entre les domaines Kerberos Windows et Linux (Figure 8) pour que toute demande d'accès à une ressource authentifiée par le KDC Linux soit relayée au KDC d'AD, qui s'assure auprès de son homologue Linux de la validité de la requête.

Au final, cette solution donne toute satisfaction. La partie la plus délicate concerne en fait, la disponibilité et la configuration de Kerberos sur les clients NFSv4.

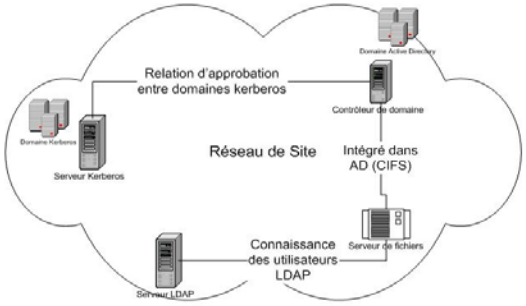

Figure 8 - *Authentification NFSv4*

5 Bilan

En 12 mois, nous avons déployé 7 serveurs NAS, une solution de sauvegarde centrale, des réplications entre NAS et une procédure de PRA. Le volume de données utilisé sur ces serveurs NAS a été multiplié par cinq depuis la mise en service.

Ce projet NAS nous a permis de renouveler les solutions de stockage de chaque site et de construire une solution mutualisée pour tout l'INRIA. Cette solution est évolutive : elle supporte l'augmentation de volumétrie, le rajout de NAS pour améliorer la disponibilité des données, sans casser l'existant. Elle constitue une infrastructure sur laquelle nous pouvons construire des services applicatifs sécurisés.

Il reste néanmoins quelques points à améliorer. Compte-tenu des caractéristiques actuelles de nos connexions inter sites, le mode travail de secours à distance et la reconstruction de volumes par le réseau WAN ne sont pas viables. Le principe d'une réplication centralisée reste toutefois valide, en tant que solution de consolidation de la sauvegarde et de reprise sur sinistre.

Sur le plan humain, la charge d'exploitation, répartie sur les membres de l'équipe, est estimée à l'équivalent d'un temps plein. Ce projet a donc permis de mutualiser réellement les ressources d'exploitation par rapport à des solutions indépendantes. Par ailleurs, ce service de sauvegarde du NAS nous a déchargé très nettement des tâches de restauration de fichiers.

L'accroissement des volumes, l'hébergement de données de plus en plus critiques (le mail, l'agenda, le Web...), les exigences croissantes en terme de sécurité et de disponibilité des données et des applications, nous amènent à consolider notre solution sur 3 axes:

- répondre aux besoins d'extension en filers et volumétrie ;
- prendre en compte la haute disponibilité, avec des solutions de cluster NAS sur les sites ;
- améliorer le PRA avec des solutions de réplication locales et les futures offres (?) de RENATER pour un service VLAN de niveau 2.

Le NAS de réplication et son système de sauvegarde devront s'adapter pour répondre à ces évolutions. Ces besoins d'extensions et d'évolutions, de nature multiple vont être couverts par la mise en place fin 2007 d'un marché à bon de commande de trois ans.

Le bilan de ce projet NAS est donc globalement positif, il nous faut toutefois adapter le PRA aux exigences attendues en terme de disponibilité. Nos prochains investissements iront en ce sens.

Bibliographie

[1] Romaric Guillier, Ludovic Hablot, Pascale Primet, Sébastien Soudan. *Rapport de recherche INRIA n° 6047*, novembre 2006.

[2] *NetApp Library*. Network Appliance, Inc, Aug 2007

[3] *Apache HTTP Server Documentation*. Apache Software Foundation, 2007.

[4] NetApp Knowledge Base. *MySQL certifications with NetApp Storage*. Network Appliance, Inc. Solution ID kb21468

[5] Eric Barrett, Bikash R. Choudhury, Bruce Clarke, Blaine McFadden, Tushar Patel, Ed Hsu, Christopher Slater. *Network Appliance™ Best Practice Guidelines for Oracle®*. Network Appliance, Inc. March 2006 | TR-3369

[6] Sanjay Gulabani. *Linux (RHEL 4) 64-Bit Performance with NFS, iSCSI, and FCP Using an Oracle® Database on NetApp Storage.* Network Appliance, Inc. Octobre 2006 | TR-3495

[7] *Manuel de référence de MySQL 5.0*, MySQL AB Août 2007

[8] NetApp Knowledge Base. *Using partitions on Linux with NetApp LUNs*. Network Appliance, Inc. March 2007 | Solution ID kb8190

[9] Peter Radkov, Li Yin, Pawan Goyal, Prasenjit Sarkarand Prashant Shenoy, *A Performance Comparison of NFS and iSCSI for IP-Networked Storage*. *3rd Usenix Conference on files and storage technologies*, March 2004.

[12] http://www.pdc.kth.se/heimdal/

[10] Sanjay Gulabani, *Linux® (RHEL 4) 64-Bit Performance with NFS, iSCSI, and FCP Using an Oracle® Database on NetApp Storage*. Network Appliance, Inc. Oct 2006 | TR-3495

[11] M. Vaughn Stewart and Michael Slisinger, *Network Appliance and VMware Virtual Infrastructure 3 Storage Best Practices*. Network Appliance, Inc. September 2007 | TR-3428

[12] S. Shepler, B. Callaghan, D. Robinson, R. Thurlow, C. Beame, M. Eisler, D. Noveck, Network *File System (NFS) version 4 Protocol*, RFC3530, April 2003