Virtualisation du stockage au DSI de l'INSERM

Julio Martins - Responsable Pôle Infrastructures
INSERM DSI
16 avenue Paul Vaillant Couturier, 94800 Villejuif

Emilie Ly - Chef de projets infrastructures
INSERM DSI
16 avenue Paul Vaillant Couturier, 94800 Villejuif

Sébastien Maury - Responsable d'exploitation développement
INSERM DSI
60 rue de Navacelles, 34394 Montpellier Cedex 5

Michel Stypak - Responsable d'exploitation production
INSERM DSI
16 avenue Paul Vaillant Couturier, 94800 Villejuif

Résumé

Le DSI de l'INSERM a lancé un projet de virtualisation du stockage en 2009.
Après un état de l'art effectué au cours du premier semestre 2009, un appel d'offres ouvert a été lancé durant l'été. La solution retenue est une des solutions phares de virtualisation du stockage : le San Volume Controller ou SVC (IBM).

La solution SVC nous a permis pour un coût maîtrisé de :

- *simplifier, centraliser et gagner en souplesse sur l'administration du stockage ;*
- *mieux sécuriser les infrastructures de stockage et la disponibilité des environnements lors des opérations de maintenance ;*
- *gagner en performance ;*
- *prendre en compte l'hétérogénéité des solutions SAN du marché ;*
- *mettre en place une solution évolutive ;*
- *mettre en œuvre des fonctionnalités avancées (réplication inter-sites, clonages d'environnements).*

L'article présentera notre démarche pour ce projet, le choix de la solution retenue, un retour d'expérience sur sa mise en production (fin 2010 début 2011) et les horizons qu'elle nous ouvre.

Mots clefs

Stockage, SAN, virtualisation du stockage, SVC, San Volume Controller

1 Introduction

1.1 L'INSERM

L'Institut National de la Santé Et de la Recherche Médicale (INSERM) est un Etablissement Public à caractère Scientifique et Technologique entièrement dédié à la santé humaine, sous la double tutelle des ministères de la recherche et de la santé. Le Département du Système d'Information (DSI) de l'INSERM héberge les infrastructures des services centraux de l'Institut et garantit la continuité des applications nationales.

Les infrastructures du système d'information sont réparties sur plusieurs sites :

- la production à Villejuif ;
- le développement au CINES (Centre Informatique National de l'Enseignement Supérieur), à Montpellier.

1.2 L'existant

Avant le lancement du projet en 2009, l'état du stockage était le suivant.

Sur chaque site le stockage SAN était constitué de 3 baies de stockage indépendantes acquises progressivement; il s'agissait de baies Sun 6140.

Sur le site de production, la volumétrie totale était de 10 To utiles FC et 25 To utiles SATA. Sur le site de développement, celle-ci était de 7 To utiles FC et 40 To utiles SATA.

Les serveurs étant connectés en double attachement fibre, l'outil de multipathing utilisé était celui fourni et supporté par le constructeur de la baie de stockage (Redundant Dual Active Controller ou rdac).

2 Le projet

2.1 Les contraintes et les objectifs

Le Pôle Infrastructures exploite plus d'une centaine d'environnements pour toutes les applications hébergées sur les différents sites. L'objectif principal était de simplifier l'exploitation et pouvoir gérer de façon souple l'espace disque attribué à ces environnements.

Les acquisitions d'infrastructures de stockage se font de manière progressive, selon les besoins, et sont régies par le code des marchés publics. Le parc de baies de stockage peut donc devenir hétérogène au cours des années en fonction des marchés successifs mis en place, ce qui amène des procédures d'exploitation différentes selon les constructeurs. Nous voulions donc également nous affranchir de ces contraintes d'hétérogénéité du parc avec une solution prenant en compte les équipements de stockage existants et les différentes solutions matérielles du marché, par souci d'indépendance par rapport à celles-ci.

Les premières baies acquises par le DSI étaient limitées en extension de volumétrie. L'ajout de volumétrie supplémentaire passait par l'extension des baies existantes (ajout de disques ou de tiroirs) ou par l'acquisition de nouvelles baies. La maintenance, l'ajout d'espace, la réorganisation de l'espace de stockage nécessitaient de ce fait des interruptions de service. Nous souhaitions à ce titre une solution évolutive dans le temps et avec un minimum d'interruptions de service lors de chacune des évolutions.

La sécurisation des équipements est primordiale ; le niveau de redondance des équipements déjà en place devait être maintenu, voire amélioré, afin de n'ajouter aucun SPOF (Single Point Of Failure).

Enfin, un dernier objectif consistait à mettre en œuvre la réplication de données entre nos deux sites via la solution de stockage.

2.2 La démarche

Un état de l'art des solutions existantes a été fait en 2009 à travers la consultation des différents constructeurs dans le domaine du stockage. C'est ainsi que nous avons monté plusieurs réunions techniques avec les constructeurs EMC, Hitachi, IBM, NetApp et Sun. Très vite est apparu que la solution répondant à toutes nos contraintes et objectifs était la virtualisation du stockage.

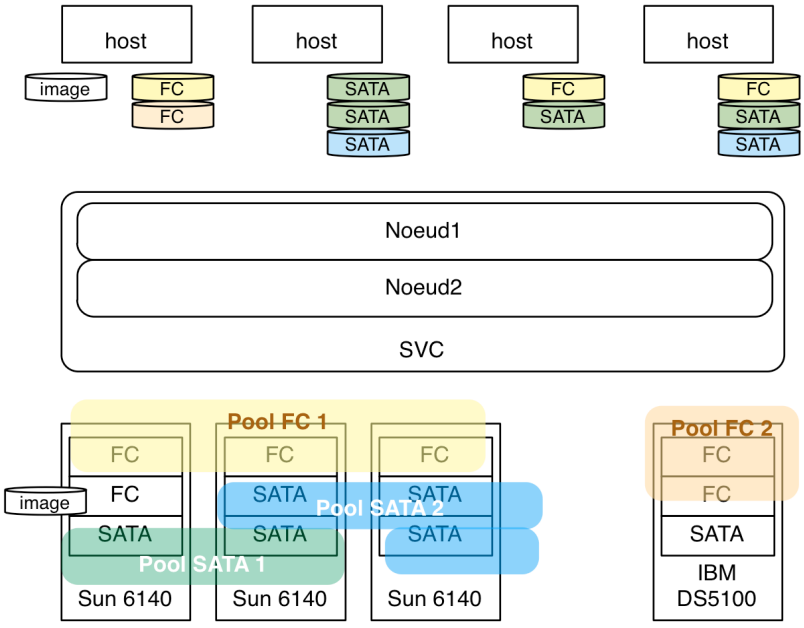
Nous avons ainsi rédigé notre expression de besoin et lancé un appel d'offres ouvert à la fin de l'été 2009 pour la mise en place d'un marché constitué de deux lots : un premier lot pour l'acquisition et la mise en œuvre d'une solution de virtualisation du stockage et un second lot pour l'acquisition de baies de stockages supplémentaires.

Quatre propositions ont été reçues ; parmi celles-ci, deux d'entre elles étaient surdimensionnées par rapport à nos besoins et par là-même très au-delà de notre enveloppe budgétaire. L'offre que nous avons retenue, dans le cadre d'un marché notifié en 2009, est celle du distributeur OVESYS nous proposant la solution SAN Volume Controller (SVC) d'IBM.

Les points forts de cette solution sont les suivants.

- Elle permet d'adresser un parc de baies de stockage hétérogène comme une seule entité de stockage auprès des serveurs et supporte plus d'une centaine de modèles de baies, pour une quinzaine de constructeurs différents.
- Le stockage natif peut être présenté en mode « image » : la LUN source est présentée telle quelle au serveur au travers du SVC ; cela permet de pouvoir faire un retour arrière éventuellement et présenter du stockage hors SVC sans avoir à migrer les données à froid.
- La forte évolutivité de la solution répond parfaitement à notre besoin, non seulement en termes de volumétrie mais également en termes de performance.
- L'outil de multipathing natif sous Linux, Device Mapper Multipath, est supporté ; celui-ci étant intégré à la distribution Linux, cela nous évite l'installation d'un outil tiers à recompiler à chaque mise à jour du noyau et le redémarrage des serveurs en conséquence. Auparavant, l'outil imposait une migration à chaque changement de baie afin de respecter les matrices de compatibilité et les conditions de support. De plus, il était différent selon le constructeur.
- L'organisation par pools de disques et la répartition des données sur l'ensemble des baies gérées optimisent les performances.

Présentation du stockage aux serveurs comme une seule entité



2.3 Mise en production

2.3.1 Préparation et installation de la solution

Après validation technique de la solution de virtualisation, il a fallu définir avec le titulaire du marché la procédure de migration en déterminant les étapes et planifiant les actions. La contrainte principale était de réduire au minimum les interruptions de service.

La préparation s'est faite au cours du premier semestre 2010 et l'installation sur le site de développement a été réalisée durant l'été 2010. La version du SVC installée était la 5.1.

Sur le site de production, l'installation de la solution a été réalisée début 2011. Nous avons pu installer la version 6 du SVC qui apporte une interface graphique plus conviviale et intuitive. De plus, cette version embarque la console d'administration au niveau des nœuds du cluster SVC. Cette version a permis de s'affranchir de l'utilisation de l'ancienne console d'administration sous Microsoft Windows, sous laquelle nous avions rencontré des dysfonctionnements (contournés par l'utilisation des lignes de commande à la place de l'interface graphique). La nouvelle interface graphique d'administration indique pour toute action les commandes, ce qui permet de scripter aisément les tâches redondantes pour les automatiser.

Le SVC du site de développement a été mis à niveau par la suite, à chaud et sans interruption de service.

2.3.2 Migration de l'existant

Pour chaque environnement la bascule du stockage natif vers le stockage virtualisé s'est faite très rapidement : le temps de désinstaller l'ancien outil de multipathing et configurer DM-Multipath, toutes les étapes préparatoires ayant été réalisées de manière anticipée (zoning, configuration), l'indisponibilité a été réduite à moins d'une heure par serveur. La présentation en mode image (présentation directe de la LUN source au travers du SVC) a grandement contribué à la réduction de l'interruption de service.

La seule exception s'est posée pour des serveurs ayant des LUN de plus de 2 To ; dans ce cas, la migration a dû être réalisée à froid car le SVC ne permet pas de gérer des volumes sources de taille supérieure à 2 To.

Par la suite, la réorganisation du stockage a pu se faire totalement à chaud, une fois les serveurs migrés, avec les fonctionnalités de clonage et de migration de volumes à chaud d'un pool de disques à l'autre.

Sur le site de développement, la migration a pu être réalisée en soirée sans interruption de service pour les utilisateurs. Le stockage sur ce site est totalement migré depuis l'été 2010.

Sur le site de production, nous avons découpé les applications à migrer par lots. Un premier lot d'applications a été migré en avril 2011 en mode image. Un second lot (concernant nos serveurs de messagerie électronique) a été migré durant l'été 2011, notamment avec des volumes de plus de 2 To que nous avons dû migrer à froid. Cela a été l'occasion de redécouper les volumes.

Il reste un dernier lot à migrer concernant notre application comptable et financière dont nous attendons une revue de l'architecture (renouvellement de serveurs, migration OS et SGBD) avant de programmer la migration du stockage vers le SVC. Ce dernier lot ne remet pas en cause l'architecture du stockage virtualisé déjà en production pour toutes les autres applications.

2.4 Caractéristiques et fonctionnalités

2.4.1 Stockage unifié

Le SVC repose sur un cluster de serveurs, appelés nœuds. La configuration et le cache sont synchronisés sur l'ensemble des nœuds du cluster, assurant la redondance en cas de défaillance d'un des nœuds.

La configuration de base avec deux nœuds couvre notre besoin actuel et est évolutive jusqu'à huit nœuds.

	Configuration à 2 noeuds	*Configuration à 8 noeuds*
Nombre de serveurs max	256	1024
Gestion de la volumétrie source	2048 x 2 To	8192 x 2 To

Notre configuration actuelle est décrite ci-dessous.

	Site de production	*Site de développement*
Nombre de serveurs connectés	14	8
Volumétrie gérée par le SVC		
Volumétrie FC gérée	6 x 1,7 To 8 x 1,8 To	4 x 1, 7 To
Volumétrie SATA gérée	16 x 1,5 To	28 x 1,5 To
Volumétrie présentée par le SVC		
Volumétrie FC utile	25 To	6,8 To
Volumétrie SATA utile	24 To	42 To

L'utilisation de l'ensemble des baies pour définir les pools de disques apporte un gain en performance :

- les disques sont organisés en pools selon leur technologie et type (vitesse, capacité) ;
- les blocs d'un même volume adressé à un serveur sont répartis sur l'ensemble des baies ; le nombre d'axes disques est multiplié, ce qui améliore grandement les performances en écriture ;
- le cache de chacun des contrôleurs des baies est utilisé de manière optimale par le SVC.

Dans nos environnements, le gain en performance a surtout été observé sur la lecture à partir de snapshots pour les sauvegardes (16 Go/heure en moyenne avant, 100 Go/heure en moyenne maintenant).

2.4.2 Sur-allocation et copies de volumes

Nous bénéficions aussi de fonctionnalités qui n'étaient pas disponibles sur les baies que nous avions acquises auparavant, notamment la sur-allocation d'espace (allocation dynamique ou thin provisioning). Un volume en allocation dynamique a une capacité réelle et une capacité virtuelle. La capacité virtuelle est celle vue par l'hôte sur lequel le volume est présenté et est supérieure à la capacité réelle du volume. La capacité réelle d'un volume peut être augmentée à chaud manuellement ou automatiquement (positionnement de seuils d'alarmes possible). Cela permet d'allouer des volumes avec une capacité virtuelle importante et d'ajuster la capacité réelle à chaud selon l'évolution des besoins. On optimise ainsi l'allocation des ressources.

Le SVC apporte des mécanismes de copies avancés avec les copies de volumes, la mise en miroir et Flash Copy.

La mise en miroir permet de copier un volume vers un pool dont la taille d'extent (unité de base du pool de stockage) est différente.

Flash Copy permet de copier simultanément plusieurs volumes source et cible, de différentes manières (basique, incrémentielle, en cascade, à plusieurs cibles).

Actuellement, nous utilisons les volumes à allocation dynamique essentiellement pour réaliser des snapshots montés pour les sauvegardes. La multitude de modes de copie de volumes va nous donner beaucoup de possibilités, notamment pour réaliser des clonages d'environnements.

2.5 Utilisation de LVM

Nous avons pu étendre l'utilisation de LVM sur les volumes pour lesquels nous réalisions des snapshots au niveau du stockage. En effet, il n'était pas aisé de réaliser des snapshots de volumes sous LVM à cause de la duplication des méta-données LVM. Aujourd'hui cela est possible grâce aux commandes de gestion des clones (vgimportclone) incluses dans certaines distributions de Linux.

Au niveau du SVC, les groupes de cohérence assurent l'intégrité de la copie de l'ensemble des volumes au sein d'un même VG, lançant simultanément la copie de chacun des volumes.

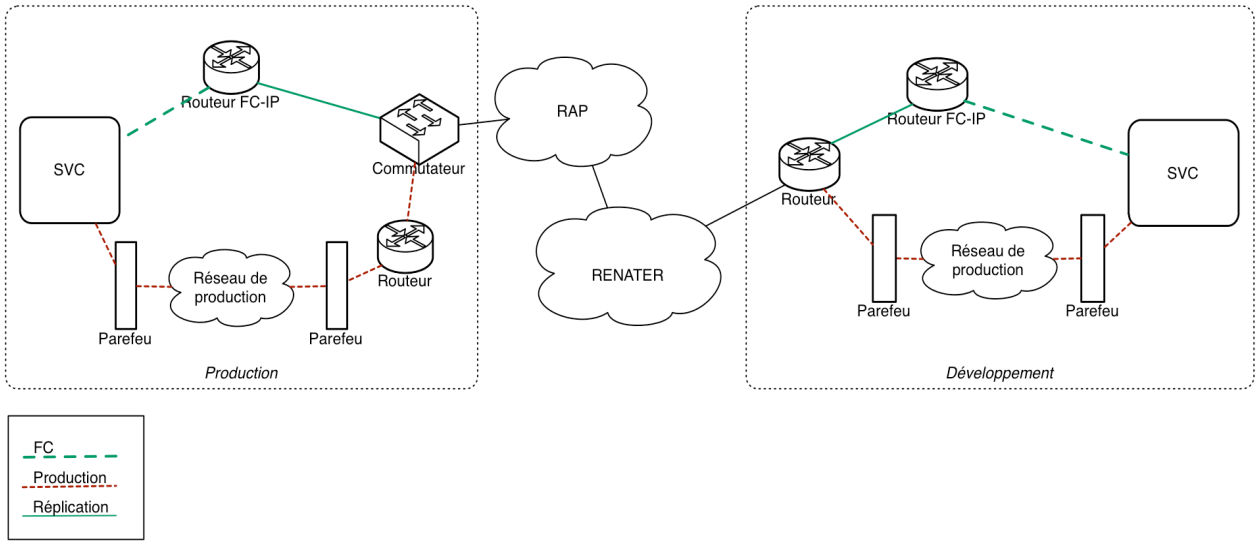
2.6 La réplication des données entre les deux sites

2.6.1 Le réseau de réplication

Du fait de la distance entre le site de production et le site de développement, la réplication à mettre en œuvre est asynchrone. Les données répliquées transitent par le réseau Ethernet. La communication entre les deux solutions de réplication est assurée par un routeur FC-IP sur chacun des sites.

Le VLAN utilisé pour la réplication est routé sur RAP (Réseau Académique Parisien) puis sur RENATER afin d'avoir un réseau privé entre nos deux sites. Ce réseau est dépourvu de chiffrement ; un tunnel VPN entre les deux routeurs FC-IP permet de sécuriser les flux de réplication qui y transitent (la sécurisation est mise en œuvre par d'autres moyens dans le cas de l'utilisation de ce réseau par d'autres applications).

Réseaux en jeu pour la réplication



Le débit de réplication est bridé à 50 Mbps par la licence en place sur les routeurs FC-IP. Ce débit semble suffisant pour une réplication asynchrone. Nous reconsidérerons l'augmentation du débit en fonction des temps de réplication constatés. Le débit du réseau est de 1 Gbps de bout en bout, ce qui nous permettra d'augmenter aisément le débit de réplication si nécessaire.

2.6.2 La réplication

La réplication se définit entre les deux solutions de virtualisation : une relation de partenariat «Global Mirror» est établie entre les deux SVCs une fois les tunnels FC-IP montés entre les deux routeurs. Chaque routeur dispose de deux ports Ethernet ; chaque port est utilisé pour monter un tunnel.

Le niveau de granularité correspond au niveau volume (LUN virtuelle) : une relation de copie à distance est créée entre le volume source d'un site et le volume cible de l'autre site. Le sens de réplication peut être aisément inversé : le volume cible peut devenir source et vice-et-versa.

Les volumes cibles de réplication peuvent également être sujets à d'autres mécanismes de copie (Flash copy).

Le tableau suivant donne les résultats des premiers tests réalisés.

Volume source (Go)	*Volumétrie allouée (Go)*	*Volumétrie réelle (Go)*	*Type de données*	*Durée*
800	700	350	Base de données	19h00
25	25	4	Ldap	2h30
576	576	350	Documents bureautique	15h00

Nous sommes toujours en phase de tests et de paramétrage afin d'optimiser les débits.

Avant de configurer la réplication des volumes, nous allons valider la bascule et la reprise des données dans les deux sens avec une application pilote. Ce projet est en cours.

2.7 La suite

La virtualisation du stockage sur chacun de nos sites nous ouvre de nouvelles perspectives. Les prochaine étapes sont les suivantes :

- la mise en place de la réplication pour les applications en production :
    - des clonages d'environnements de production à Villejuif vers le site de Montpellier pourront être réalisés pour des besoins de développements, tests et recettes ;
    - la politique de sauvegarde pourra être revue, notamment sur les externalisations ;
    - le site de développement pourra être intégré dans le cadre d'un Plan de Reprise d'Activité ;
- le test des fonctionnalités iSCSI ;
- la mise en place de stockage partagé pour les hyperviseurs ESX sur SVC ;
    - la mise à jour du SVC en version 6.2 qui intègre le support des APIs Vmware vStorage (VAAI, ou vStorage APIs for Array Integration) en est un pré-requis.

3 Conclusion

Comme dans tout projet de cette envergure, nous nous sommes heurtés à quelques difficultés depuis la mise en production du SVC.

Le premier point est la mise à jour des composants, que ce soit du SVC et/ou des différents composants SAN (baies, commutateurs, routeurs). La version de firmware de nos baies Sun 6140 n'était plus dans la matrice d'interopérabilité de la dernière version du SVC. Nous avons émis une demande de RPQ (Request Product Qualification) et sommes en attente du résultat.

Nous avons eu une autre surprise concernant la prise en compte des licences de réplication. En effet, à la notification du marché fin 2009 était prise en compte la volumétrie source. Récemment, le constructeur a changé de politique sur la prise en compte de ces licences : la volumétrie cible doit également être prise en compte. Des négociations sont en cours afin de trouver un accord.

Malgré ces difficultés, la solution mise en place pour la virtualisation du stockage apporte des avantages notoires et incontestables. Elle nous a permis d'atteindre les objectifs de centralisation et de souplesse d'administration, avec en prime un gain en performance important.