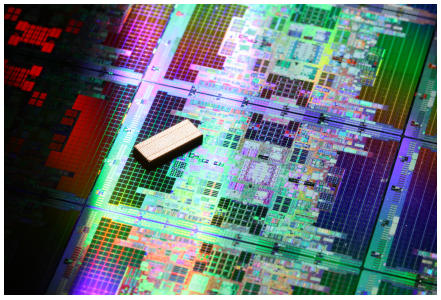
Pierre-Luc Delisle
Stockage de données en réseau
9 Février 2014

Le stockage de fichiers mis en réseau

Essai

“The desktop metaphor was invented because one, you were a stand-alone device, and two, you had to manage your own storage. That's a very big thing in a desktop world. And that may go away. You may not have to manage your own storage. You may not store much before too long.”

- Steve Jobs



Déni de responsabilité (*Disclaimer*)

Cet essai a été écrit suite à une recherche personnelle sur les technologies de stockages. Ce document peut refléter mes opinions personnelles. Il fut rédigé aux meilleures de mes connaissances.

Introduction

Le stockage de données est, de nos jours, une division de l'informatique qui est de plus en plus en vogue et qui ne cesse d'être en perpétuelle transformation. Le stockage de données a toujours été un secteur important de l'informatique dans lequel les contraintes sont excessivement grandes. En effet, les solutions de stockages se doivent aujourd'hui d'être incroyablement rapides et efficaces, tout en restant fiables. Il y a cependant des contraintes. Je note trois contraintes majeures au niveau du stockage de fichiers informatiques :

- La densité
- L'évolutivité
- La consommation électrique

En effet, que ce soit chez un particulier ou une entreprise, l'entité aura toujours de plus en plus de fichiers à stocker. William Cain, actuel vice-président des technologies chez Western Digital, affirme que selon des experts analystes, il y aura en 2016 en moyenne 3.3 TO de données par ménage. Idéalement, pour des raisons de simplicité, l'utilisateur ou l'administrateur désire ne pas multiplier les unités de stockages abusivement. La densité du stockage s'avère donc une contrainte majeure dans le but de garder un système de stockage simple, fiable et efficace, tout en gardant une grande capacité de stockage. En date de l'écriture de ce document, nous pouvons atteindre une densité maximale de 6 TO de stockage par disque dur, ce qui représente tout de même une énorme quantité de données pour un seul disque. Actuellement, les fabricants de disques durs peuvent fabriquer des plateaux de 750 gigabits par pouce. Dans un futur rapproché, Western Digital avec la technologie HARM, ou Heat Assisted Magnetic Recording créée par Seagate, prévoit plus que quadrupler la densité de leur plateau, augmentant la densité à 4 térabits par pouce. De ce fait, Seagate prévoit construire des disques durs ayant une capacité de 20 TO de donnée d'ici 2020, et des disques durs pour ordinateurs portables de 10 TO dans le même laps de temps.

L'évolutivité de la solution de stockage entre par la suite en considération. Personne ne veut véritablement une solution qui le limitera dans un futur plus ou moins éloigné et encore moins

lorsque l'on parle de stockage de données. Il est important de bien doser son budget afin de s'offrir une solution ayant un bon rapport prix/évolutivité, car bien sûr, cette caractéristique a également son prix.

Finalement, la consommation électrique a également son mot à dire dans les solutions de stockage. La consommation électrique vient principalement de deux composantes dans le système :

- Les disques durs en grand nombre
- Le processeur, carte-mère et contrôleur de disques

Bien sûr, les disques durs consomment de l'énergie électrique. Toutefois, il faut noter les formidables avancées technologiques qui ont permis de réduire considérablement la consommation au cours des dernières années. On peut dorénavant avoir un disque dur de qualité *enterprise grade data center* qui a une vitesse de rotation de 7200 tours/minute consommant moins de 10w en fonctionnement. Toutefois, multipliez cette valeur par une solution de stockage à grande échelle, la consommation électrique n'a d'autre choix que d'augmenter de façon linéaire au nombre de disques durs. C'est pourquoi un disque dur de plus grande densité permet un meilleur ratio capacité/consommation électrique.

Vient ensuite la mise en réseau de ce stockage pour partager les données avec d'autres utilisateurs du réseau local. Plusieurs technologies existent et il y a également plusieurs moyens de mettre en réseau une unité de stockage. Il est important de bien comprendre et de démystifier ces méthodes. Ce document en présentera quelques-unes.

Ayant pris en considération ces facteurs et contraintes, il sera plus facile de trouver une solution adaptée à l'environnement de travail.

La problématique et présentation du travail

Récemment, je me suis retrouvé avec la problématique de stockage la plus commune qui soit : le manque de capacité. Avec présentement deux batteries RAID 1 totalisant 2.5 TO de capacité de stockage, les statistiques me décrivaient un taux d'utilisation à plus de 90 %. J'ai alors commencé à étudier les propos qui se retrouveront dans cet essai, essayant de déterminer la solution la plus appropriée à mon usage.

Veuillez prendre note qu'aucune configuration logicielle ne sera présentée dans ce document. Toutefois, à l'issue de ce document, je serai en mesure d'éclaircir l'achat d'une solution

de stockage pour moi-même et qui s'appliquerait également dans une petite entreprise ou tout autre individu visant le même usage que moi. Je documenterai toutes les étapes de conception et de mise en place d'un laboratoire de virtualisation et de la mise en place de mon réseau dans des documents subséquents. Le document ci-présent met l'emphase sur la présentation de diverses technologies et solutions de stockage ainsi que les avantages et désavantages de chacune d'elles.

Les besoins

Toute bonne étude de cas repose sur l'établissement des besoins et de l'utilisation, tout en gardant une perspective à long terme.

Pour bien cerner les besoins, je me suis établi un questionnaire auquel j'ai répondu le plus adéquatement possible.

Le questionnaire

1. Pourquoi déployer une nouvelle solution de stockage ?
   Car je serai très rapidement en manque d'espace de stockage. De plus, je voudrais bâtir un laboratoire de virtualisation à des fins d'expérimentation et d'apprentissage.

2. Quelle sera l'étendue de la solution de stockage (nombre d'utilisateurs, sites, charge de travail - *Workload*).
   Le nombre d'utilisateurs actifs sera restreint : 2 ou 3 utilisateurs. Le nombre d'utilisateurs potentiel se situe à environ 10. Les sites varient dans la province de Québec. Chaque site aura un utilisateur. La charge de travaille sera relativement lourde puisqu'il y aura de la virtualisation parmi les utilisations.

3. Quels sont les buts ?

   - Stockage mis en réseau
   - Avoir plusieurs services incluant :
     - Serveur VPN
     - Serveur DNS local / Serveur DHCP
     - Serveur de fichiers FTP et serveur web HTTP

    - Serveur multimédia
- Média de stockage pour machines virtuelles Microsoft Hyper-V.
- Stockage de sauvegardes Windows, Linux et Apple Time Machine.

4. Quelle serait la capacité de stockage requise à moyen terme ?
   Environ 6 TO utilisables.

5. Est-ce que l'évolutivité est importante ?
   Oui.

6. Quelles sont les attentes en performances ?
   Assez performant pour servir de média de stockage pour plusieurs machines virtuelles opérant en même temps.

7. Quel est le niveau de sécurité requis ?
   Protection contre un bris de disque dur + sauvegarde externe (*One hard disk drive failure + external backup*).

Les possibilités

Parmi les contraintes mentionnées dans la section précédente, un autre facteur n'a pas encore été discuté et sera l'objet principalement traité dans cet essai : les performances.

Il existe plusieurs types de solutions offrant chacune des performances différentes avec des niveaux de sécurité différents. Parmi celles-ci, je décrirai les principales solutions que j'ai étudiées puisqu'elles se sont avérées intéressantes pour mon usage :

- DAS, ou *Direct-Attached Storage*
- NAS, ou *Network Attached Storage*
- SAN, ou *Storage Area Network*
- Stockage avec système d'exploitation virtualisé - Serveur ESXi

Chacune de ces possibilités offre son lot d'avantages et d'inconvénients. J'essayerai de démystifier chaque lacune et chaque bénéfice de ces technologies, tout en les appliquant à mes besoins actuels et futurs.

Mes buts futurs

Après mûre réflexion, j'ai déterminé que mes occupations et buts futurs génèreraient beaucoup de données. En effet, je compte faire un lot de certifications techniques et voir même entamer des études universitaires. Parmi ces certifications, les certifications en réseautique *CCNA Security, Data Center, Voice, Wireless,* avec la possibilité de faire les certifications *CCNP*. Les certifications dans le domaine du *Cloud Computing* m'intéressent également. Les certifications Microsoft et VMware sont toutefois deux certifications que je ferai assurément et qui, à elles seules, requièrent énormément d'espace disque en raison du parc de machines virtuelles à construire au long de ces certifications.

Toutes ces applications génèreront beaucoup de données. Il ne faut pas oublier les sauvegardes périodiques de ces systèmes ainsi les *snapshots* des machines virtuelles.

Il me faudra donc un système de stockage fiable, rapide, ne nécessitant pas de maintenance tout en ayant une capacité de stockage extensible.

Explications des solutions de stockage et des principes réseau

DAS - *Direct-Attached Storage*

Un DAS est un système de stockage directement attaché au serveur. Les données transigent sur le serveur ou une station de travail localement, sans passer par un réseau quelconque. Un système DAS est principalement un boitier avec plusieurs disques durs interconnectés par un *Host Bus Adapter,* ou HBA. Un HBA est simplement un contrôleur de bus très rapide passant par exemple du bus PCI-Express au bus SATA via un contrôleur logique. Les protocoles utilisés dans les DAS sont principalement le ATA, SATA, eSATA, SCSI, SAS ou Fiber Channel.

Cette solution est centralisée sur le côté serveur et est de loin la plus rapide puisqu'elle exploite des bus de communication très rapides.

La solution DAS fut, jusqu'à aujourd'hui, ma solution de stockage utilisée. Toutefois, avec plusieurs serveurs sur le réseau, cette solution de stockage en fait une solution décentralisée et plus difficile à gérer. Puisque ce sont généralement des serveurs ou des stations de travail complètes, ils génèrent beaucoup plus de chaleur et consomment plus d'énergie qu'un système

de stockage centralisé et dédié uniquement au stockage. De plus, un DAS est généralement complètement isolé du réseau (sinon, il devient un NAS). Un DAS ne peut donc pas partager de données avec d'autres systèmes DAS ou de machines clientes.

Le système DAS était, jusqu'à présent, ma solution de stockage employée puisque tous les disques durs étaient centralisés sur ma machine principale, sans mise en réseau.

NAS - *Network Attached Storage*

Un NAS est un dispositif de stockage mis en réseau et qui donne accès à des données au sein d'un groupe de clients hétérogène[1]. Un NAS peut faire tourner un service de partage de fichiers, mais également une multitude d'autres services. Il est cependant spécialisé pour partager et servir des clients en fichiers. Son matériel et sa configuration logicielle sont en majeure partie dédiés à cet usage. Un NAS peut être vendu comme étant un ordinateur spécialisé comme serveur de fichiers.

Depuis 2010, les NAS gagnent beaucoup en popularité à cause de leur facilité de déploiement ainsi que la multitude de méthodes de partage de fichiers qu'ils intègrent. Ils peuvent également faciliter grandement l'administration, en plus de centraliser le stockage des données.

Très souvent, les NAS sont équipés de plusieurs disques durs arrangés en batteries RAID et disposent d'une redondance en cas de panne d'un ou de plusieurs disques durs. Ils évitent la multiplication des serveurs de fichiers sur le réseau et disposent souvent de composantes à plus faible consommation d'énergie. Ils utilisent une variété de protocoles de partage tels que le NFS, SMB/CIFS ou AFP.

SAN - *Storage Area Network*

Un SAN est un réseau dédié au stockage de fichiers qui donne accès aux blocs de stockages consolidés (*block level consolidated data storage*). Il s'agit d'un réseau spécialisé

---

[1] **Heterogeneous computing** systems : réfère à un système électronique qui utilise une variété d'unités computationnelles de types différents. Une unité computationnelle peut être un general-purpose processor (GPP), un special-purpose processor (par exemple un digital signal processor (DSP) ou un graphics processing unit (GPU)), un co-processor, ou custom acceleration logic (application-specific integrated circuit (ASIC) ou field-programmable gate array (FPGA)). En général, une plateforme informatique hétérogène consiste en un processeur avec différentes architectures d'instructions (instruction set architectures (ISAs)).
http://en.wikipedia.org/wiki/Heterogeneous_computing

permettant de mutualiser des ressources de stockages. L'expression *block level data storage* est en fait un accès bas niveau aux disques durs de stockage.

Dans un SAN, les ressources de stockages apparaissent comme étant directement attachées aux serveurs. Un SAN a son propre réseau de stockage qui n'est normalement pas accessible depuis le réseau local par les autres équipements réseau (machines clientes).

Un SAN fait abstraction des fichiers; il n'a aucun système de fichier. Un SAN ne fait que des opérations de bas niveau sur les disques durs. Toutefois, il existe des systèmes des fichiers bâtis pour des SAN qui eux un accès aux fichiers.

Essentiellement, un SAN converti plusieurs îlots DAS ayant du stockage directement connecté en les consolidant ensemble au sein d'une même batterie de disques dans un réseau à très haute vitesse (Fiber Channel).

Une autre différence serait dans l'apparence des disques durs sur la machine serveur. Sur un SAN, les disques apparaissent comme étant directement connectés, alors que dans le cas d'un NAS, il s'agit de dossiers de partage sur un serveur de fichiers.

Dans un SAN, la plupart des réseaux utilisent le protocole de communication SCSI entre les serveurs et les disques durs eux-mêmes. Une couche de mappage à d'autres protocoles est utilisée pour former un réseau :

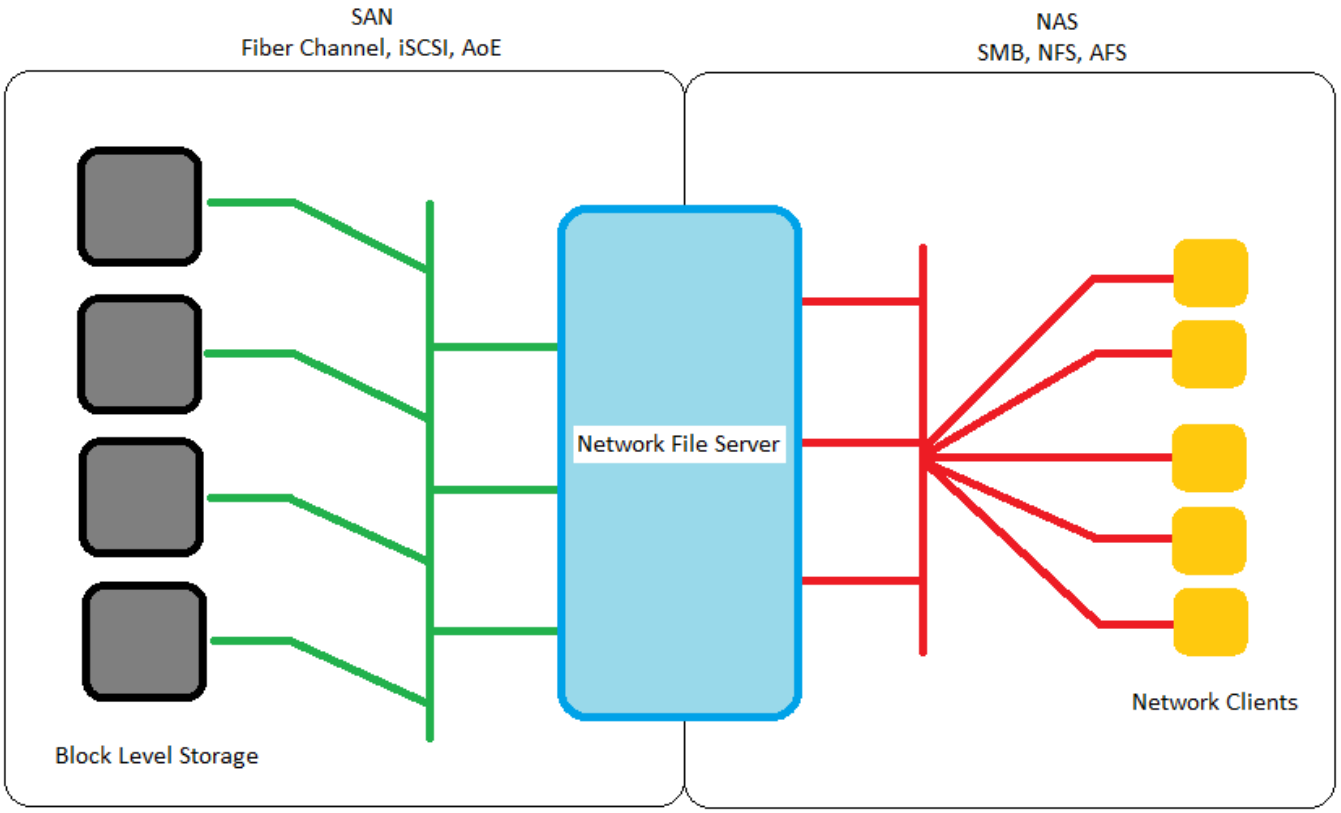
- ATA over Ethernet (AoE)
- Fibre Channel Protocol (FCP)
- Fibre Channel over Ethernet (FCoE)
- ESCON over Fibre Channel (FICON) (utilisé pour les mainframe)
- HyperSCSI, mapping of SCSI sur Ethernet
- iFCPor SANoIP, mappage du FCP sur IP
- iSCSI, mappage du SCSI sur TCP/IP
- iSCSI Extensions for RDMA (iSER), mappage du iSCSI sur InfiniBand

Notez qu'un SAN peut également être mis en place avec un *QoS*, ou qualité de service, pour améliorer les performances du réseau en cas de forte utilisation.

Un autre fait est à prendre en note. Un SAN est habituellement mis en place pour fournir du stockage à des serveurs de virtualisation tournant sur Microsoft Hyper-V, VMware ESXi ou encore les solutions Citrix. Le stockage apparaît alors comme étant local au serveur. Le protocole iSCSI offre normalement dans bien des cas de meilleures performances que le SMB ou NFS pour ce genre d'utilisation.

Un SAN est pourvu de bien des avantages. Parmi ceux-ci, l'architecture SAN permet de simplifier l'administration et ajoute plus de flexibilité au stockage. Les serveurs peuvent également démarrer directement du SAN (*boot from SAN*). Cela permet de remplacer rapidement des serveurs fautifs ou mal configurés. Ils ont également tendance à être plus efficaces lors de recouvrement après un désastre (*disaster recovery*). On peut également configurer un caching du côté serveur et ainsi augmenter drastiquement la capacité de I/O par seconde. Plusieurs solutions permettent également de cloner à la volée et de faire des captures instantanées (*snapshots*) du disque consolidé.

Cependant, un SAN peu s'avérer très coûteux en raison de l'équipement utilisé. En effet, des commutateurs à fibre optique coûtent encore très cher aujourd'hui. On peut toutefois concevoir un SAN sur des connexions Ethernet conventionnelles (câble en cuivre), mais un tel réseau sacrifie beaucoup les performances du SAN. La norme Ethernet 10 Gbps devient



Différences entre un SAN et un NAS

aujourd'hui un moyen économique pour implanter un SAN sans sacrifier les performances, mais reste globalement moins rapide que de la fibre optique.

Serveur de virtualisation ESXi

La virtualisation est un sujet qu'un essai à lui seul ne pourrait couvrir entièrement. Je m'efforcerai ici de n'offrir qu'une définition de ce qu'est un *hypervisor* (hyperviseur en français), puisque la prise de connaissance de la virtualisation devrait être un prérequis à la lecture de ce document.

Un serveur de virtualisation ESXi est un serveur ayant comme "système d'exploitation" un hyperviseur. Un hyperviseur ne peut être utilisé seul. Une fois installé puis configuré, on doit, par le biais d'un client, lui créer et lui mettre en place les machines virtuelles qu'il pourra opérer.

Il existe deux types d'hyperviseur :

1. Type 1 Bare-Metal Hypervisor
2. Type 2 Hosted Hypervisor

Dans le premier cas, il s'agit d'un système qui n'opère que des machines virtuelles. Le système en l'occurrence VMware ESXi ou encore Microsoft Hyper-V, ne sert qu'à virtualiser des machines *guests*. Un hyperviseur, comme celui de VMware, est un système d'environ 1 million de lignes de codes seulement (contrairement à un système d'exploitation qui peut en contenir facilement 25 millions de lignes). Cela procure toujours dans le cas de VMware ESXi, une surface d'attaque de loin réduite à celle d'un véritable système d'exploitation. Il en résulte un système monolithique[2] réduit en taille et qui procure une plus grande sécurité. Le cas de Hyper-V est légèrement différent puisqu'il opère conjointement avec le système d'exploitation Windows, mais l'hyperviseur est essentiellement un service distinct qui n'emprunte l'environnement Windows que pour le GUI. Les hyperviseurs fonctionnent conjointement de très près avec les technologies embarquées sur le matériel, notamment la VT-x et la VT-d (et les équivalents AMD).

Dans le deuxième cas, on parle de virtualisation logicielle. On passe par un hyperviseur sous forme logicielle, par le biais d'un système d'exploitation Windows, Linux ou Mac OS X. Les

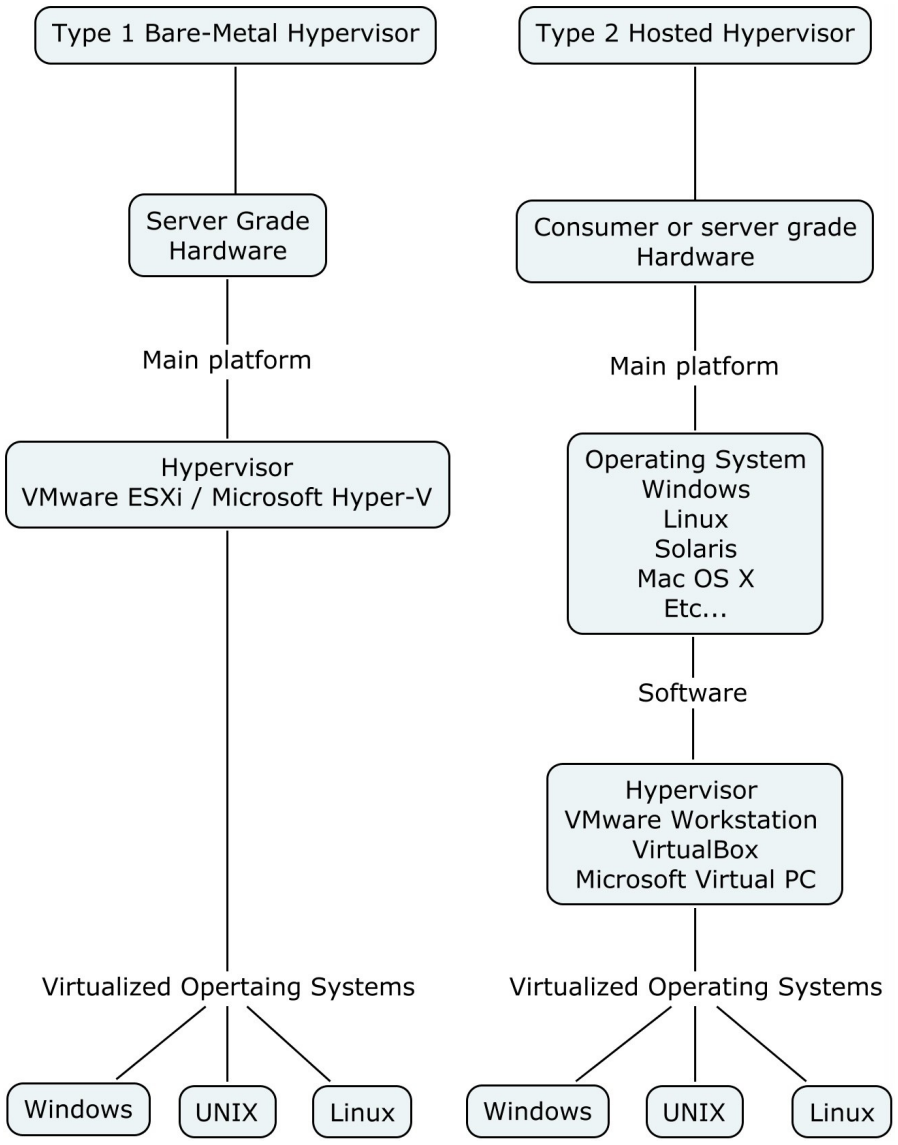
---

[2] Un **noyau monolithique** est une architecture de système d'exploitation dans laquelle le kernel (noyau) est un processus qui travaille dans un espace mémoire défini. Le kernel est seul dans cette espace et fonctionne en mode supervision. Le modèle monolithique se distingue des autres architectures de systèmes d'exploitations (tels que l'architecture micro-noyau) car elle définit une interface virtuelle de haut niveau sur le matériel informatique. Un ensemble d'appels de système (*calls*) mettent en œuvre tous les services du système d'exploitation telles que la gestion des processus et la gestion de la mémoire. Le kernel peut invoquer des fonctions directement. Il est statique et non modifiable, mais des pilotes de périphériques peuvent être ajoutés au noyau sous forme de modules.

performances sont moindres que dans le premier cas puisqu'il y a une couche applicative entre le matériel et l'hyperviseur. Hyper-V, bien qu'intégré au sein de Windows, n'est pas un hyperviseur de type 2. Il n'opère pas par dessus Windows, mais bien juste au-dessus du matériel.

La beauté d'un hyperviseur de type 1 réside dans le fait que l'on peut virtualiser n'importe quel système d'exploitation, y compris des systèmes d'exploitation de stockage. Dans le dernier cas, on peut très bien faire opérer ce genre de système tout en faisant tourner d'autres machines virtuelles sur d'autres plateformes. Cela permet de réduire les coûts puisque toutes les plateformes peuvent opérer sur le même serveur de virtualisation. Dans cette solution, le stockage n'est pas virtualisé; seul le système d'exploitation (par exemple, FreeNAS) est virtualisé et a accès aux disques durs physiques grâce au *pass-through* des contrôleurs logiques (LSI ou Intel) de la carte mère via la technologie VT-d ou AMD-Vi.

Vous trouverez ci-dessous une image comparative des hyperviseurs.



LUN - *Logical Unit Number*

Un LUN est un numéro d'identification pour identifier une unité logique qui est adressée par le protocole SCSI ou un protocole qui encapsule le protocole SCSI tel que les protocoles précédemment énuméré dans la présentation d'un SAN. Il s'agit d'une nomenclature de disque logique.

802.3ad - *Dynamic Link Aggregation*

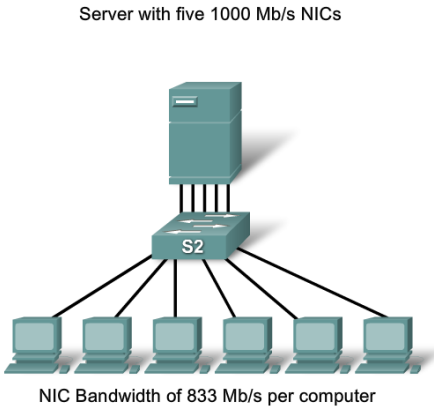
La notion de *Link Aggregation* - ou LAG - est très importante puisqu'elle joue un très grand rôle au niveau des performances d'un réseau. Chez Cisco, c'est le protocole LACP 802.3ad - *Link Aggregation Control Protocol* - qui est pris en charge.

Le *Link Aggregation* est, dans sa définition la plus simple, un moyen de multiplier les liens entre des serveurs et des équipements réseau de façon à augmenter la capacité théorique de la bande passante. Une agrégation de liens permet également une certaine protection en cas de panne avec la possibilité d'implanter une redondance (*failover link*).

Un LAG est effectué dans la partie inférieure du modèle OSI et peut être implanté soit dans la couche physique, liaison de données ou la couche réseau. Le LAG qui nous intéresse dans le cas d'un réseau de petite entreprise est celui qui s'effectue dans la deuxième couche du modèle OSI et qui agrège des liens entre des serveurs et un commutateur.

Pour former un LAG, quelques prérequis sont nécessaires. Tout d'abord, il faut des liens *full duplex point to point.* Ces liens doivent également avoir le même débit (*data rate*).

Physiquement, pour former un LAG, on doit installer deux câbles ou plus entre un serveur ou une station de travail et un commutateur. On doit ensuite réaliser de façon logicielle le «*LAN Teaming*» qui consiste à créer une «équipe» avec les cartes réseau présentes dans les machines. Ce *LAN Teaming* doit se souscrire à la norme 802.3ad pour former, avec le commutateur, un lien d'agrégation dynamique. Notez qu'on peut également configurer un LAG entre des commutateurs pour augmenter la bande passante entre les équipements d'un réseau local. Cisco désigne l'appellation de EtherChannel pour ce type d'agrégation de liens. L'image ci-contre exprime bien la topologie de ce type de lien.



Dans le cadre de l'agrégation de bande passante, on doit déterminer si un commutateur à agréger dispose d'assez de ports pour prendre en charge la bande passante requise. Prenons l'exemple d'un port Gigabit Ethernet, qui peut traiter jusqu'à 1 Gbit/s de trafic. Si vous disposez d'un commutateur à 24 ports et que tous les ports peuvent être exécutés à des débits en gigabits, vous pouvez générer jusqu'à 24 Gbit/s de trafic réseau. Si le commutateur est connecté au reste du réseau à l'aide d'un seul câble réseau, il peut uniquement transférer 1 Gbit/s de donné vers le reste du réseau. En raison de l'engorgement au niveau de la bande passante, les données sont transférées plus lentement. 1/24e du débit du câble est alors disponible pour chacun des 23 périphériques connectés au commutateur. Le débit du câble décrit le débit de transmission de données maximum théorique d'une connexion. Par exemple, le débit du câble d'une connexion Ethernet dépend des propriétés physiques et électriques du câble, associées à la couche la plus lente des protocoles de connexion.

L'agrégation de liaisons aide à réduire les goulots d'étranglement de trafic en permettant d'associer jusqu'à huit ports de commutateur pour les communications de données. Ainsi, un débit de données pouvant atteindre 8 Gbit/s est obtenu lors de l'utilisation de ports Gigabit Ethernet. Cela empêche par le fait même de connecter d'autre équipement terminal au commutateur puisque plus de ports sont employés à des fins d'*uplink.* Avec l'ajout de plusieurs liaisons ascendantes 10 Gigabit Ethernet (10GbE) sur certains commutateurs Cisco, des débits très élevés peuvent être atteints.[3]

Toutefois, il y a des limitations à cette technologie. Certes, le 802.3ad augmente la bande passante théorique disponible entre des équipements réseau. Cependant, une seule conversation IP est possible sur un lien d'agrégation. Cela n'implique donc pas que deux machines peuvent communiquer à 2 Gbps et plus s'ils ont tous deux des liens d'agrégation entre eux. Les communications entre les deux dispositifs réseau demeureront à 1 Gbps.

Un lien d'agrégation sert alors à multiplier les liens de plusieurs machines vers un serveur donné, ou encore de multiplier les liens entre des commutateurs afin d'augmenter la capacité de transfert entre l'équipement. Toutefois, bien qu'on multiplie les liens physiques, la vitesse de communication entre deux points reste de 1 Gbps. Le trafic réseau est seulement divisé par le nombre de liens entre les équipements.

Cela est dû au fait que le protocole TCP/IP présent sur le réseau local afin de délivrer les paquets de façon fiable entre les machines. Le protocole TCP ajoute une surcharge au niveau de la communication afin d'assurer la livraison des paquets. Il doit attendre l'accusé de réception du destinataire.

---

[3] Le passage provient des textes de la certification *CCNA - Routing and Switching,* module 3, et a été adapté à l'écriture de cet essai.

On peut cependant configurer les réglages de l'équilibrage de charge (*load balancing*) de l'EtherChannel. Plusieurs types d'équilibrage de charge peuvent être configurés sur un commutateur Cisco 2960-S :

Type	Description
dst-ip	Équilibrage de charge basée sur l'adresse IP du serveur de destination.
dst-mac	Équilibrage de charge basée sur l'adresse MAC du serveur de destination du paquet entrant.
src-dst-ip	Équilibrage de charge basée sur l'adresse IP serveur source et de destination.
src-dst-mac	Équilibrage de charge basée sur l'adresse MAC source et de destination.
src-ip	Équilibrage de charge basée sur l'adresse MAC source et de destination.
src-mac	Équilibrage de charge basée sur l'adresse MAC source du paquet entrant.

Notez que l'équilibrage de charge est actif globalement sur le commutateur. On ne peut donc pas configurer plusieurs types d'équilibrage de charge sur un commutateur ou un type par Port Channel.

En conclusion, un LAG est une très bonne pratique à effectuer dans les réseaux où un serveur est fortement sollicité, ou encore pour éviter les goulots d'étranglement (*bottleneck*). C'est également une fonction relativement facile à implémenter. Toutefois, il ne procure pas N fois la bande passante entre deux points donnés (où N représente le nombre de liens physiques). Il requiert également un peu plus de gestion de la part de l'administrateur. Finalement, un LAG est bien, mais ne sera jamais équivalent à un lien de base plus rapide, comme un lien Infiniband 40 Gbps ou un lien Ethernet 10 Gbps, qui acheminent l'information beaucoup plus rapidement sur un seul lien au lieu de le distribuer comme le fait l'EtherChannel.

MPIO - *Multi-Path I/O*

À la base, le *multi-path I/O* est une technique utilisée pour bénéficier de meilleures performances ou d'une certaine redondance (*fault-tolerance*) en ayant plus d'un chemin physique entre le processeur *host* et le stockage de masse en passant par différents bus, contrôleurs et/ou commutateurs, tout en interconnectant ceux-ci.

Le MPIO est en quelques points une technologie semblable au 802.3ad *Dynamic Link Aggregation* en termes de caractéristiques. Les avantages d'une telle mise en place sont multiples :

- Dynamic load balancing
- Traffic shaping
- Automatic path management
- Dynamic reconfiguration

J'ai décidé d'en parler dans mon essai dans le cadre de la virtualisation, où le MPIO trouve généralement sa place avec les divers hyperviseurs tels que Microsoft Hyper-V et VMware ESXi.

Une grande différence existe cependant entre un LAG 802.3ad et le MPIO. Le LAG est effectif dans un NAS, alors que le MPIO est effectif dans un SAN. De ce fait, les deux technologies ne fonctionnent pas du tout de la même façon. Comme discuté précédemment, le 803.ad ne permet pas d'accélérer le débit entre deux points d'un même flux d'entrées/sorties. Une opération d'entrée/sortie traversera toujours un seul lien (*path*). Le 802.3ad ne procurera pas de meilleures performances pour le protocole iSCSI, il ne peut que procurer une redondance en cas de défaillance et un *load balancing* des opérations.

Le MPIO fonctionne quant à lui entre un initiateur et un *host*. La première utilité du MPIO est de concevoir un réseau de haute disponibilité (*high availability*) pour former une redondance du chemin employé vers le média en cas de panne. Il existe cependant d'autres façons de configurer le MPIO, dont plusieurs gèrent l'équilibrage de charge :

- Fail Over Only : politique de contrôle qui n'utilise qu'un seul *path* et laisse les autres à titre de redondance en cas de panne.
- Round Robin : équilibrage de charge entre les différents *paths* de manière à augmenter les performances globales.

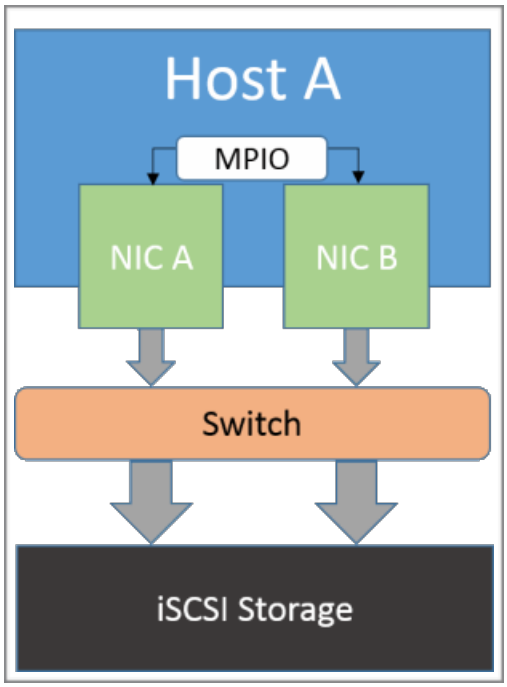
- Round Robin with Subset : politique d'équilibrage de charge qui accorde à l'application de spécifier combien de liens sont disponibles en équilibrage de charge et combien d'autres sont disponibles à titre de liaisons de redondance.

- Least Queue Depth : politique d'équilibrage de charge qui permet d'envoyer des opérations d'entrées et sorties sur un *path* avec le plus petit nombre d'*outstanding I/O* [4].

- Weighted Paths : politique d'équilibrage de charge qui accorde une priorité relative à chaque *path*.

- Least Block : politique d'équilibrage de charge qui envoie des opérations d'entrées/sorties sur le *path* qui possède le moins de blocs de données actuellement en cours de traitement.

Avec tous ces procédés d'équilibrage de charge, le MPIO peut utiliser efficacement deux ou plusieurs interfaces Ethernet ensemble, de façon à augmenter le débit des données. C'est un moyen d'utiliser plusieurs interfaces Ethernet de façon simultanée afin d'améliorer le débit de l'information y circulant.



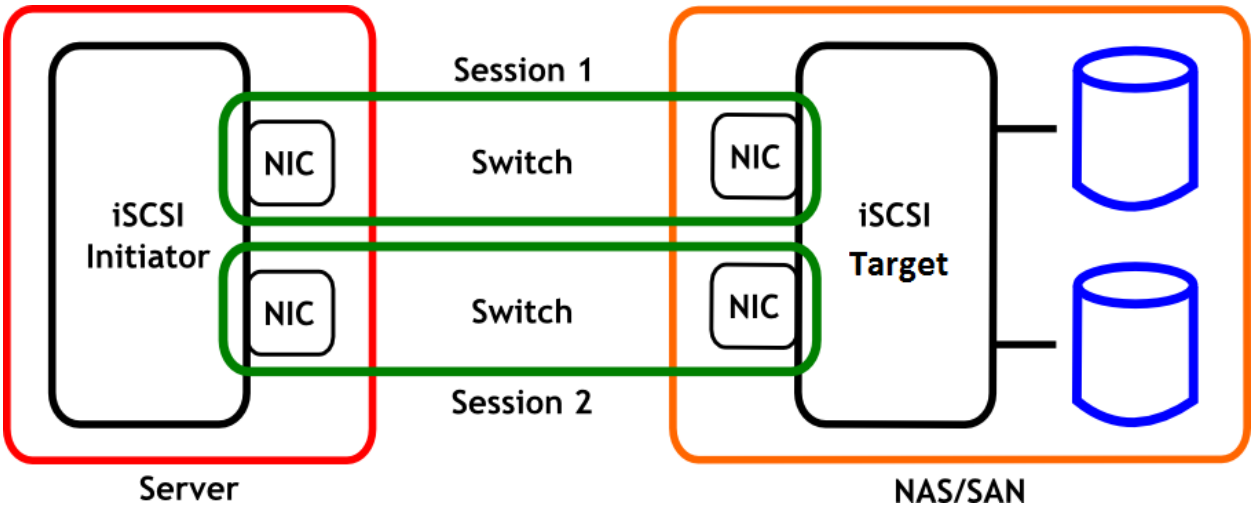
Le MPIO est logiciel, et non pas matériel. Il faut alors configurer correctement le système d'exploitation en vue de faire fonctionner cette technique, car le pilote MPIO ne peut fonctionner correctement et efficacement sans avoir découvert, énuméré et configuré les différents périphériques de stockage. Le système d'exploitation verra ces périphériques à travers un groupe logique d'adaptateurs en redondance. Le système comprendra que plusieurs chemins mènent au même système de stockage via plusieurs liens physiques.

La technique MPIO permet de prévenir la corruption des données en s'assurant que la gestion du pilote associée au système de stockage unique distant (par exemple, la carte réseau) à travers plusieurs liens physiques est correcte. Dans le cas où le MPIO n'est pas utilisé, cela peut entraîner des pertes de données puisque le système d'exploitation croit qu'il a affaire à deux unités

[4] Un *Outstanding I/O* est une opération de lecture/écriture mis en file d'attente par le contrôleur. Ce dernier peut les remettre en ordre, réduisant le temps entre chaque I/O pour un traitement plus rapide. C'est une profondeur de file d'attente.

de stockages distants reliés par deux chemins différents. Il ne fait donc aucun traitement de données en série (*data serialization*) et ne prévient pas les conflits dans la cache (*caching conflicts*).

Sous forme visuelle, le MPIO en redondance pourrait prendre cette forme :



RAID - *Redundant Array of Independent Disks*

La technologie RAID fait partie des technologies de stockage les plus utilisées au monde. Elle permet une chose capitale : la redondance de disques et la sécurité des données.

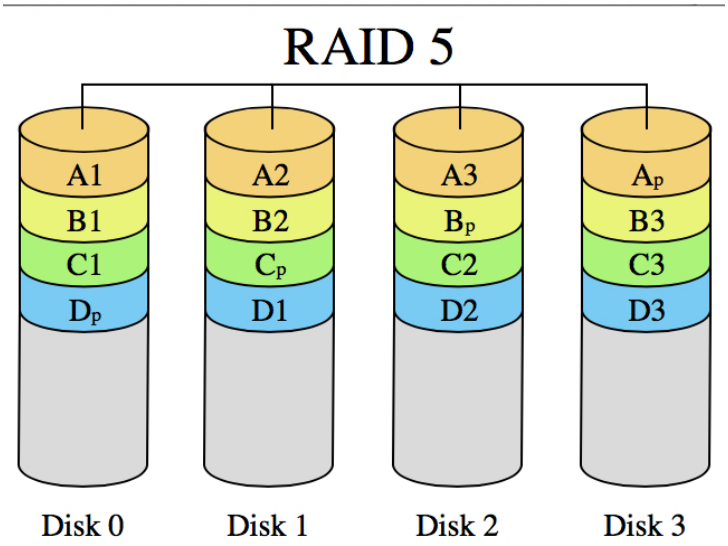
En effet, le stockage est probablement le domaine de l'informatique qui est le plus à risque. Personne, et encore moins une entreprise ne souhaite voir disparaître ses données en raison d'une défaillance physique du matériel. Or, le disque dur étant la seule pièce mécanique d'un ordinateur avec ses têtes de lecture/écriture et ses plateaux en rotation constante, il n'est pas rare qu'un disque dur présente des signes de défaillances ou arrête subitement de fonctionner en raison d'un bris mécanique interne après quelques années d'utilisation 24/7. Lorsqu'un disque dur brise mécaniquement, il entraîne malheureusement avec lui les données qu'il contenait.

C'est ici que le RAID vient à la rescousse des données en faisant de la redondance. Les disques durs sont regroupés sous forme de batteries (*array*). Il existe plusieurs niveaux de RAID et chacun offre un certain niveau de sécurité. Voici une liste des principaux types de RAID :

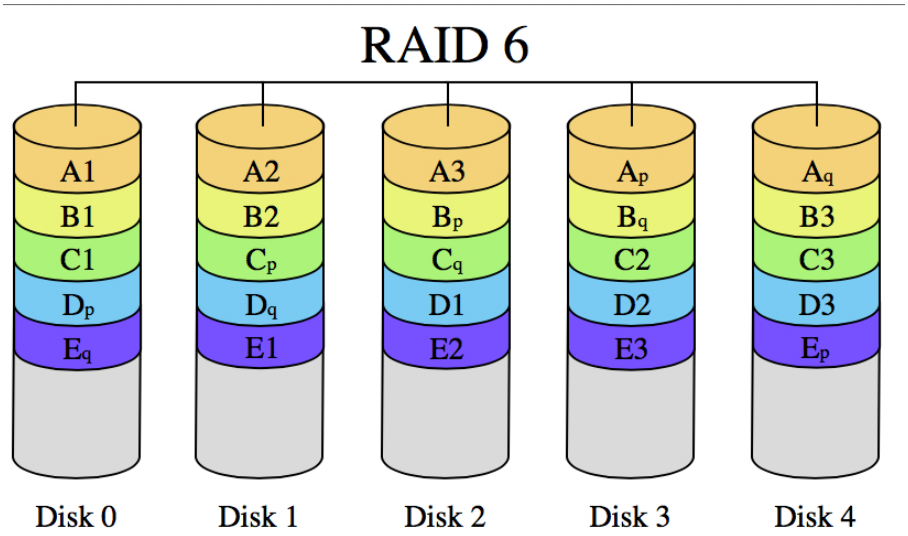
- **RAID 0** : *block-level stripping without parity or mirroring*. Niveau de RAID qui ne dispose d'aucune redondance ou tolérance de faute, mais améliore les performances avec un parallélisme entre les disques durs lors des opérations d'écriture et de lecture. Un bris d'un disque dans la batterie RAID entraîne la perte de toutes les données.

- **RAID 1** : *block-level mirroring without parity or stripping*. Les données sont écrites de manière systématique et identique sur tous les disques durs de la batterie. Il offre un niveau de protection de *n-1* disque(s). Lors d'un bris de disque dur, les données sont toujours accessibles tant et aussi longtemps qu'un disque de la batterie fonctionne.

- **RAID 5** : *block-level stripping with distributed parity*. Tout comme le RAID 0, les données sont divisées à travers tous les disques de la batterie, entraînant de ce fait des améliorations en terme de performances. Le RAID 5 est pourvu d'une redondance distribuée à travers tous les disques de la batterie. Il procure une tolérance de faute d'un disque dur. Si deux disques brisent en même temps, les données de la batterie sont perdues puisque le contrôleur ne peut plus calculer la parité manquante. La batterie doit contenir un minimum de trois disques.



- **RAID 6** : *Block-level stripping with double distributed parity.* Tout comme le RAID 5, le RAID 6 procure un parallélisme au travers des disques durs de la batterie RAID, mais effectue une double parité sur deux disques durs en même temps. Cela lui procure donc une tolérance de panne de deux disques durs avant de perdre toutes les données de la batterie. La batterie doit contenir un minimum de 4 disques.



Il existe également des combinaisons de ces niveaux de RAID, notamment le RAID 10 ou encore le RAID 0+1.

Généralement, le RAID est contrôlé par un contrôleur physique, communément appelé “carte RAID”. Parmi les meilleurs du marché, on note Intel, LSI et Areca. Ces cartes ont toute la logique nécessaire pour contrôler des dizaines, voir centaines de disques durs. Malheureusement, ces cartes sont coûteuses, mais également très performantes.

Cette solution représente toutefois un certain désavantage. Si elle décharge le processeur *host* principal de la charge de calcul pour les opérations I/O et de calcul de parité, les disques durs sont totalement dépendants de ce contrôleur. Si ce dernier cesse de fonctionner pour quelconque raison, il faut retrouver un contrôleur identique ou qui partage les mêmes propriétés physiques (même contrôleur embarqué) pour récupérer les données sur la batterie RAID. De plus, une batterie doit être ajoutée à la carte. En effet, ces dernières sont très souvent pourvues d’une mémoire vive DDR2/3 à des fins de caching d’opérations I/O et de caching de données. En cas de coupure de courant, la mémoire vive ne retient pas les données. Cela peut très vite entraîner la perte de données ainsi qu’une défaillance au niveau de l’intégrité de la batterie RAID, mettant en péril non pas seulement les données stockées temporairement dans la mémoire vive, mais également celles de toute la batterie. Ces batteries entraînent un coût de maintenance supplémentaire.

Pour pallier à ces problèmes, il existe le RAID logiciel, de plus en plus en vogue. Le RAID logiciel intervient au niveau du système d’exploitation. Sur Windows, le tout est géré par la console de gestion de l’ordinateur (*Windows Computer Management Console*), alors que l’utilitaire *mdadm* (*Multiple Device Administration*) existe sur Linux. Le RAID logiciel a tendance à être beaucoup plus flexible. En effet, la batterie RAID n’est pas dépendante du contrôleur puisque le RAID est géré par le système d’exploitation. On peut donc “déplacer” la batterie RAID dans une autre machine et remonter la batterie de manière totalement logicielle et avoir accès aux données en un temps record. Le RAID logiciel permet également de limiter les coûts par rapport au RAID matériel en écartant l’achat d’un contrôleur dispendieux.

Toutefois, bien que plus flexible, le RAID logiciel a également ses lacunes. On ne peut, par exemple faire du *hot-swap* avec les disques durs, cette fonctionnalité étant réservée principalement aux contrôleurs physiques. Les performances sont également moindres que lorsqu’on acquiert un contrôleur dédié : c’est le processeur *host* qui calcul toutes les opérations nécessaires au stockage, en plus de celles des applications et kernel du système d’exploitation. Les performances varient donc selon l’usage. De ce fait, la reconstruction d’une batterie à l’état *degraded* est plus long sur un RAID logiciel que matériel.

Il y a toutefois une chose à retenir. Même si l’on parle de redondance et de tolérance de panne, le RAID ***n’est pas*** une solution de sauvegarde. Le RAID permet d’avoir une disponibilité des données (*uptime*) accrue en cas de panne matérielle, mais ce n’est pas un moyen de

sauvegarde. Un moyen de sauvegarde est entièrement logiciel avec, la plupart du temps, un *versionning* des fichiers. *Apple Time Machine* ou *Windows Backup and Restore* sont des utilitaires de sauvegarde embarqués au sein du système d'exploitation. On peut également faire une sauvegarde périodique manuelle des fichiers importants vers un disque dur externe que l'on débranche par la suite, le rendant hors ligne. Le RAID ne permet pas de sauvegarder périodiquement vers un autre média de stockage. Le RAID ne protège pas contre l'effacement ou l'écrasement accidentel de données. C'est pourquoi il ne constitue pas une méthode de sauvegarde efficace.

En conclusion :

Type de RAID	Usage	Sauvegarde
RAID Logiciel	Solution à prix plus abordable Conçu pour un seul serveur / station de travail Parfait pour des petites entreprises ou pour une solution résidentielle Très flexible	N'est pas un moyen efficace de sauvegarder des données.
RAID Matériel	*Mission Critical Usage* Usages haute performance Demande beaucoup de IOPS (par exemple, une basse de données) Solution à grande échelles, plusieurs serveurs, etc.	

Les systèmes de fichiers

On ne peut parler de stockage de fichiers sans parler de systèmes de fichiers. Plusieurs systèmes de fichiers existent présentement sur le marché, tels que le NTFS, BTRFS, EXT4, XFS, HFS+, exFAT et le ZFS pour ne nommer qu'eux. Chacun a ses forces et faiblesses. Dans le cadre de cet essai, je traiterai de quatre systèmes de fichiers, le EXT4 Linux, le XFS, le ZFS de Solaris et le BTRFS.

EXT4 - Le système de fichiers Linux

Le système de fichier EXT4 - ou *fourth extended file system* - est l'évolution du système de fichier EXT3. Le EXT4 fut introduit en octobre 2008 dans sa version stable. Il est utilisé par toutes les grandes distributions Linux. Il est le système de fichier officiel de Red Hat Enterprise Linux 6. On peut créer un volume allant jusqu'à 16 TB en EXT4. Un volume de 1 Exbibyte (2^{60} octets) est possible, mais un volume aussi gros n'est pas recommandé pour un usage en production. Parmi les avantages du EXT4, on note :

- extent-based metadata
- delayed allocation
- journal check-summing
- large storage support

L'extent-based metadata est une façon plus efficace et compacte pour mapper l'espace utilisée sur le disque. L'allocation retardée (*delayed allocation*) permet d'améliorer les performances disques globales. En effet, elle permet au système de fichier de remettre à plus tard la sélection permanente de l'emplacement où les données de l'utilisateur doivent être écrites. Cela permet au système de prendre de meilleures décisions pour écrire les données de façon plus contiguë sans fraguementation. De plus, la réparation du système de fichier (fsck) est très rapide et permet de récupérer ou d'éviter la perte de données.

On notera toutefois que le EXT4 n'a pas beaucoup d'autres options quant à la sécurité et intégrité du système de fichier. En effet, il n'y a pas de déduplication, ce procédé de compression qui élimine les doublons de fichiers. Il n'y a également pas de cryptage transparent au sein même du système de fichier, ni même de compression à la volée. Aucun *checksum* ou parité n'est établi sur le disque ; le fichier est écrit sur le plateau du disque dur sans vérification de l'intégrité du fichier. Le EXT4 n'a pas non plus la possibilité de faire des *snapshots*, une fonction qui permet de faire une copie en lecture-seule d'une donnée gelée en un point donné dans le temps. Il n'en demeure pas moins un système de fichier performant utilisé sur un très grand nombre de serveurs Linux.

XFS

Le XFS est un système de fichier qui fut introduit sur le marché en 1994 par Sillicon Graphics Inc. Il fut porté sur le kernel Linux. Le XFS a su dans son histoire se tailler une place de choix dans les systèmes et les serveurs à grande échelle (*cluster*). Cette place lui fut accordée en raison de son impressionnante performance en traitement de fichiers en parallèle, une caractéristique très recherchée dans les serveurs.

Le XFS partage quelques caractéristiques du EXT4, notamment l'allocation retardée. Il supporte également nombre de caractéristiques dignes de serveurs à grande échelle :

- delayed allocation
- dynamically allocated inodes
- b-tree indexing for scalability of free space management
- ability to support a large number of concurrent operations
- extensive run-time metadata consistency checking
- sophisticated metadata read-ahead algorithms
- tightly integrated backup and restore utilities
- online defragmentation
- online filesystem growing
- comprehensive diagnostics capabilities
- scalable and fast repair utilities
- optimizations for streaming video workloads

Ces nombreuses caractéristiques ont permis au XFS de gagner une place importante chez Red Hat, et normalement, tout ce qui rentre chez Red Hat est fiable, sécuritaire et performant. Sur le dernier point, le XFS ne manque pas et est classé parmi les plus performants du marché. Toutefois, il relâche son plein potentiel sur des batteries de stockages très larges : plusieurs dizaines voir centaines de téraoctets. Pour que le XFS performe bien, il doit également être sur des serveurs avec une bande passante très grande; plus de 200 MO/s sont requis pour exploiter le plein potentiel, ainsi que plus de 1000 IOPS. Comme dit plus haut, les applications utilisées doivent

être capables d'écrire et de lire des données sur plusieurs threads de façon à traiter les données en parallèle plutôt qu'en série, où le EXT4 performe mieux.

En conclusion, le XFS est un bon système de fichier, qui requiert certes des connaissances plus poussées que le EXT4, mais sans être trop complexe à mettre en oeuvre et à exploiter. Il se marie bien avec de très gros serveurs ayant de larges batteries de stockage et de l'équipement de communication ayant une très grande bande passante comme le Fiber Channel, 10 Gbps ou l'Infiniband, ce qui est peu probable en dehors d'un usage en entreprise.

ZFS - "*The final word to file system"*

Le ZFS fut introduit en 2005 par Sun Microsystem (Oracle par la suite). Il est très souvent qualifié de système de fichier "ultime" puisqu'il intègre pratiquement le plus grand nombre de caractéristiques possibles qu'un système de fichier peut contenir. Il est également qualifié comme étant un système de fichier "exotic" de par sa complexité inouïe. Je ne ferai ici qu'un survole de ce système de fichier puisqu'on pourrait dédier à lui seul un mémoire de maîtrise.

Le ZFS est présentement supporté sur Linux (avec un kernel modifié), OpenIndiana, FreeBSD, Illumos et Solaris. Un port sur la plateforme Mac OS X a déjà été fait, mais le projet est tombé après quelque temps. Le projet fut racheté par un ancien ingénieur d'Apple, puis revendu à GreenBytes qui supporte actuellement la seule implémentation du ZFS sur le système d'exploitation Mac OS X. Toutefois, dû aux limitations matérielles des plateformes d'Apple, le système n'a pas vraiment sa place sur la plateforme de Cupertino.

Le ZFS est un système de fichier pouvant s'étendre jusqu'à 256 Zebibytes (2^{78} octets). Le ZFS n'est pas seulement un système de fichiers, c'est également un gestionnaire de volumes logiques. Le ZFS inclut une protection active contre la corruption de donnée silencieuse (*silent data corruption*), gère la compression et le cryptage des données à la volée, *snapshots*, copie sur écriture, la déduplication, de même qu'un module d'analyse des données en continu.

Toutes ces fonctionnalités regroupées en un seul et même système en font un ensemble très complexe à gérer, mais qui, à cause de sa sécurité absolue, en fait désormais un système de plus en plus convoité en entreprise où l'on traite de l'information sensible.

Le ZFS est un système qui met beaucoup l'accent sur l'intégrité des fichiers. Il en fait même sa priorité. Il fut conçu pour résister et protéger l'utilisateur (ou plutôt l'administrateur) contre la corruption silencieuse, celle qui se produit lorsqu'il y a des pics de courant dans le disque dur, sous la forme du phénomène de *bit rot* (la dégradation des données sur un média de stockage

avec le temps sans que rien ne se passe), ou encore lorsqu'il y a des bogues dans le micrologiciel du disque dur. Elle peut également survenir lorsque les opérations de lecture et d'écriture sont mal dirigées, des erreurs de pilotes (les données se volatilisent dans un mauvais buffer dans le kernel) et finalement les écrasements de fichiers. Certes, la corruption silencieuse est assez rare, mais peut être tout de même bien réelle. Selon une étude à grande échelle, réalisée par Bairavasundaram, L., Goodson, G., Schroeder, B., Arpaci-Dusseau, A. C., Arpaci-Dusseau intitulée *An analysis of data corruption in the storage stack* dans *Proceedings of 6th Usenix Conference on File and Storage Technologies* a démontré que la corruption silencieuse est bien réelle. Sur 41 mois et couvrant plus de 1.5×10^6 disques durs, 4×10^5 incidents de corruption silencieuse fut détecté. Le CERN s'est également penché sur la question. L'organisation produisant une quantité phénoménale de données et en est arrivée à la même conclusion. La corruption silencieuse est bien réelle sur les disques durs et peut être grandement néfaste.

Pour remédier à ce problème, le système de fichier ZFS balaie continuellement la surface du disque à la recherche d'erreurs potentielles afin de les corriger. L'intégrité des données est atteinte grâce à un calcul d'un *checksum* SHA-256 dans l'arbre du système de fichier. Chaque bloc de données subit un calcul de parité et ce calcul est stocké dans le pointeur de ce bloc. Par la suite, le pointeur du bloc subit lui-même un calcul de parité qui lui est stocké dans son propre pointeur. Lorsqu'un fichier est lu, son *checksum* est calculé puis comparé avec la valeur préalablement calculée. Si la valeur n'est pas identique, le ZFS peut tenter la réparation du fichier s'il y a présence de miroir sur le pool de stockage.

Comme mentionné précédemment, le ZFS gère la copie sur écriture. Le *copy-on-write* est une méthode d'enregistrement de fichier qui, plutôt d'écraser un fichier lorsque celui-ci est réenregistré, alloue un nouveau bloc de données sur le disque dur pour écrire le fichier. De ce fait, si l'application qui a demandé l'écriture du fichier commet une erreur et *crash*, le système n'écrasera pas le fichier précédent et aucune donnée ne sera corrompue.

Mais le ZFS n'est pas seulement un système de fichiers. C'est également un contrôleur de volumes logiques. Le ZFS implémente un RAID logiciel. C'est d'ailleurs dans ce mode que le ZFS performe le mieux. Un RAID matériel peut également être envisageable, mais le système ne sera pas toujours en mesure de réparer la corruption de données, car le contrôleur RAID interférera avec les procédés du système. C'est pourquoi, lorsque l'on veut récupérer des contrôleurs RAID lors de la migration vers un système ZFS, on *flash* le micrologiciel (*firmware*) des contrôleurs RAID en mode IT qui désactive toute forme de RAID. Le ZFS gère quatre niveaux de RAID, tous similaires du RAID conventionnel :

- RAID-Z1 est l'équivalent du RAID 5
- RAID-Z2 est l'équivalent du RAID 6

- RAID-Z3 qui permet jusqu'à 3 pannes de disques durs sur la même batterie
- Mirroring est l'équivalent du RAID 1

Toutefois, il y a un énorme désavantage au ZFS : les performances. Qui dit RAID logiciel dit une énorme demande en ressources matérielles pour opérer. Le système de fichier contient plusieurs niveaux de calculs de parité. Ces calculs sont des algorithmes directement calculés par le processeur lui-même qui n'intègre pas de jeu d'instruction spécifique à ces algorithmes. Cela demande alors une énorme puissance de calcul de la part du processeur *host* comparativement aux systèmes de fichiers conventionnels. Puisque les données voyagent par la mémoire vive et y sont stockées en cache, le serveur se doit impérativement d'avoir - en grande quantité - de la mémoire RAM ECC, ce qui fait exploser les coûts de mise en place. Puisque le serveur peut être fortement sollicité, de la RAM ECC *Fully Buffered* ou *Fully Registered* peut également contribuer à une meilleure stabilité du système. Conjuguez le tout avec une très grande capacité de stockage pour gérer les *snapshots* et la redondance et vous aurez droit à un système de stockage certes des plus sécuritaire qui soit, mais également très dispendieux.

Heureusement, pour pallier légèrement à ce problème, le ZFS dispose de différentes couches de *caching* pour améliorer les performances en lecture et écriture. Idéalement, pour tourner à plein régime, les données devraient être littéralement stockées dans la RAM, mais cela est impossible pour des raisons de coûts. La RAM est toutefois utilisée en premier lieu comme cache pour les données fréquemment utilisées. Un deuxième niveau de cache associé à des SSDs est par la suite utilisé pour y stocker des données un peu moins utilisées. Finalement, les données peu utilisées sont stockées sur les disques durs, le médium le plus lent.

ARC est l'algorithme utilisé pour mettre en cache des données dans la RAM. Certains disent qu'il faut d'énormes capacités de mémoire vive pour opérer un serveur de stockage sous le système de fichier ZFS. Cela n'est pas à proprement parler une vérité. L'algorithme utilisé est très efficace et permet d'exploiter au maximum les ressources systèmes. Toutefois, il est recommandé d'avoir entre 8 et 16 GB de RAM ECC pour un serveur de "petite taille", soit de quelques dizaines de téra-octets en capacité de stockage. Si on extrapole la quantité de stockage, on peut voir que la quantité de RAM nécessaire peut grimper assez vite.

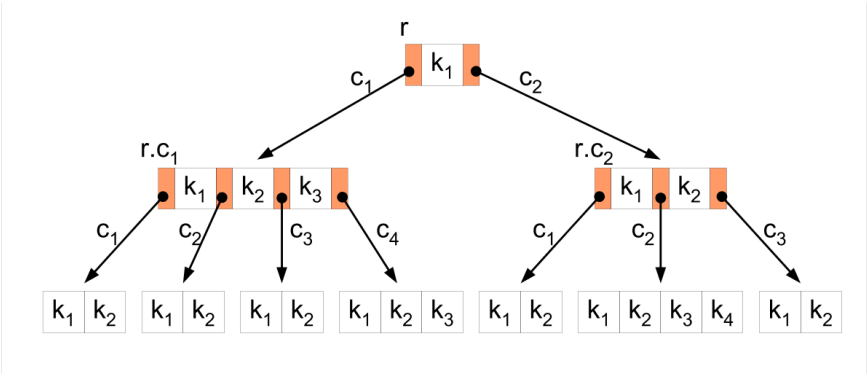
L2ARC est associé à la cache en lecture sur SSDs. L2ARC améliore considérablement la vitesse de la déduplication. ZIL est l'algorithme utilisé pour l'écriture sur SSDs. Il transforme les écritures synchrones en écritures asynchrones.

Pour l'instant, seul Oracle Solaris dans sa version 11.1 gère la dernière version du système de fichier.

En conclusion, le ZFS a bel et bien le dernier mot sur les systèmes de fichiers. Il s'agit ici d'un des systèmes de fichier des plus complexes qui soit, mais également des plus complets. Il s'agit du plus sécuritaire d'entre tous grâce à ses nombreuses caractéristiques. Une grande communauté d'utilisateurs l'utilise.

BTRFS - *B-Tree File System*

Le BTRFS est un autre système de fichier créé par la firme Oracle et fortement commandité par Red Hat Enterprise Linux. C'est un des systèmes de fichiers des plus jeunes : le développement du BTRFS a commencé en 2007. De ce fait, il n'est toujours pas certifié comme étant stable. Il est cependant distribué dans bon nombre de distributions de Linux sous forme de système de fichier expérimental à des fins de développement.



Le BTRFS fut conçu pour palier aux lacunes des systèmes de fichiers actuellement disponibles sur Linux, notamment le manque de *pooling*, *snapshots* et le *data checksumming*. De plus, la plupart des systèmes de fichiers Linux font face à des défis de taille lorsque vient le temps de les appliquer dans des applications de plus grandes étendues, comme c'est le cas aujourd'hui dans des grands centres de données.

Le BTRFS repose sur une architecture *B-Tree*. Le *B-Tree* est une architecture de données structurées en arbre. Cette structure stocke les données sous une forme ordonnée et permet une exécution des opérations d'insertion et de suppression en temps logarithmique.

En plus des caractéristiques mentionnées ci-dessus, le BTRFS dispose des particularités suivantes :

- Mostly self-healing in some configurations due to the nature of copy on write
- Online defragmentation
- Online volume growth and shrinking
- Online block device addition and removal

- Online balancing (movement of objects between block devices to balance load)
- Offline filesystem check
- Online data scrubbing for finding errors and automatically fixing them for files with redundant copies
- RAID 0, RAID 1, RAID 5, RAID 6 and RAID 10
- Subvolumes (one or more separately mountable filesystem roots within each disk partition)
- Transparent compression (zlib and LZO)
- Snapshots (read-only or copy-on-write clones of subvolumes)
- File cloning (copy-on-write on individual files, or byte ranges thereof)
- Checksums on data and metadata
- In-place conversion (with rollback) from ext3/4 to Btrfs
- File system seeding (Btrfs on read-only storage used as a copy-on-write backing for a writeable Btrfs)
- Block discard support (reclaims space on some virtualized setups and improves wear leveling on SSDs with TRIM)
- Send/receive (saving diffs between snapshots to a binary stream)
- Hierarchical per-subvolume quotas
- Out-of-band data deduplication (requires userspace tools)

Les fonctionnalités planifiées dans le développement :

- In-band data deduplication
- Online filesystem check
- Very fast offline filesystem check
- Object-level RAID 0, RAID 1, and RAID 10
- Incremental backup
- Ability to handle swap files and swap partitions
- Encryption

Comme on peut le remarquer, plusieurs de ces caractéristiques se retrouvent également chez le ZFS. Une différence demeure toutefois. Le ZFS est plutôt orienté vers les distributions exotiques (OpenIndianna, Illumos), Solaris et sur BSD Unix. Il existe certes un *port* (migration) sur Linux, en particulier sur Debian, mais cela requiert un kernel de source tierce. Or, le BTRFS se voudrait exclusif à l'architecture Linux et aux distributions de la branche Red Hat et Debian, peu importe la distribution. Ce système de fichier serait greffé au kernel Linux.

Toutefois, on ne sait pas quand le BTRFS sera disponible en version stable. Toutefois, le kernel 3.13 apporte de grandes améliorations dans le système de fichiers et on note que celui-ci a fait un pas de plus vers la sortie officielle. Cependant, toutes les pages consultées dans cette recherche mentionnent que le BTFRS est sous développement intensif pour livrer au plus vite ce nouveau système de fichiers aux entreprises et serveurs sous Linux. Lorsqu'il sortira, Linux connaîtra la plus grande révolution qu'il aura connue depuis fort longtemps. En attendant, mieux vaut rester sur le EXT4 pour stocker des données sur Linux.

Les aspects étudiés pour chaque solution de stockage

Dans chaque étude de cas, les neuf aspects suivants seront étudiés :

1. Temps de mise en place
2. Maintenance
3. Fonctionnalités
4. Configuration matérielle et spécifications
5. Évolutivité
6. Rendement énergétique
7. Performances
8. Possibilités futures
9. Sécurité

Études de cas - Cas #1 : NAS Synology

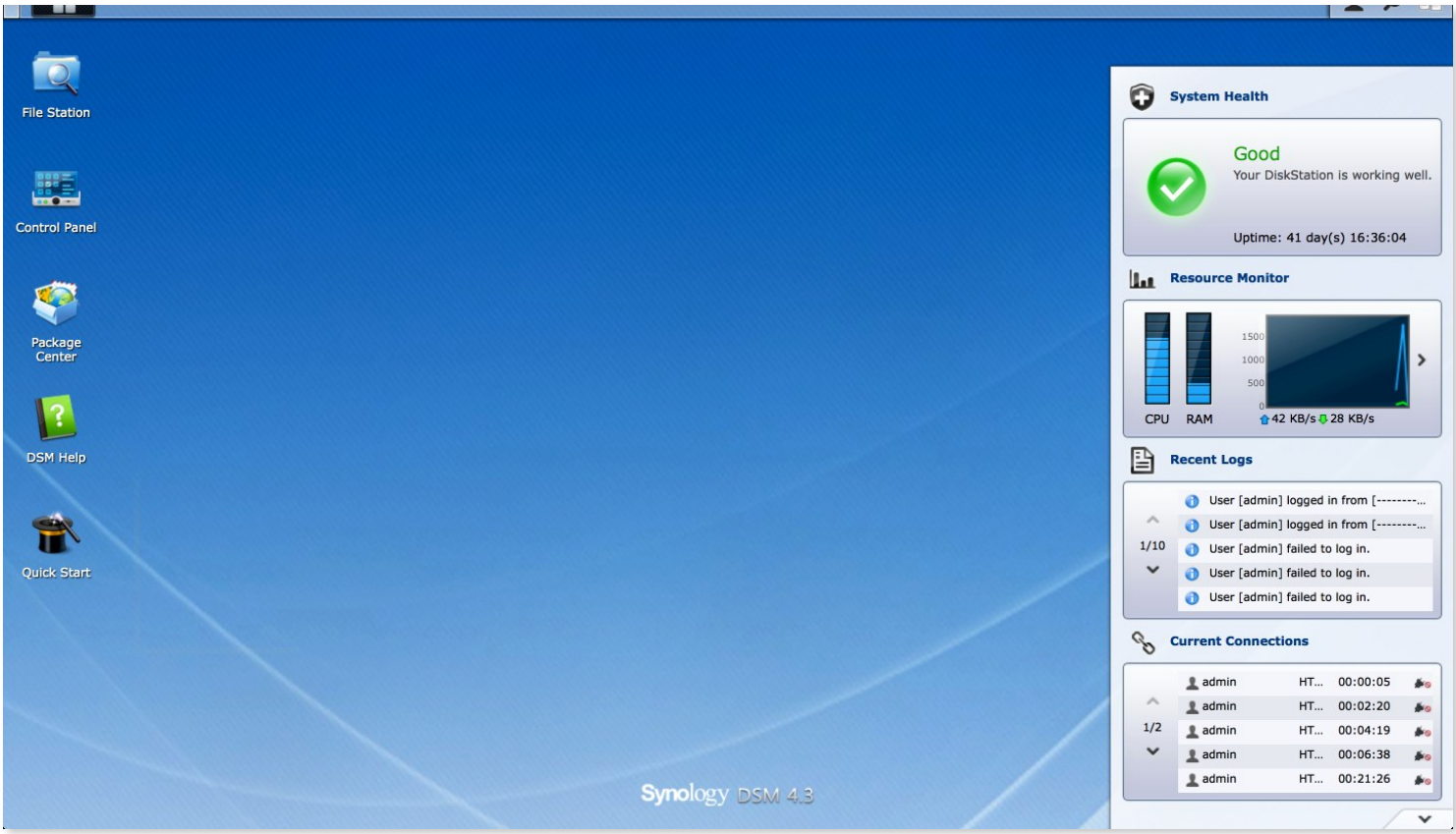
Synology est une compagnie de renommée mondiale basée à Taipei, en Taiwan et fondée par deux ex-employés de Microsoft. Synology offre plusieurs solutions de stockage, du plus bas de gamme au plus haut de gamme, en passant par des solutions *box* jusqu'au *rackmount*. Cependant, que l'ont prennent le plus bas de gamme ou le serveur de stockage le plus dispendieux, toutes les solutions s'articulent autour d'un même et unique système d'exploitation, le Synology DSM - ou DiskStation Manager -, actuellement en version 4.3 lors de l'écriture de ce document.

Synology DSM est une distribution basée sur le kernel Linux. Les unités plus bas de gamme qui sont pourvues d'un processeur ARM utilisent également le kernel Linux compilé en ARM. Synology offre également une gamme d'applications mobiles pour faciliter l'intégration d'un nuage personnel.

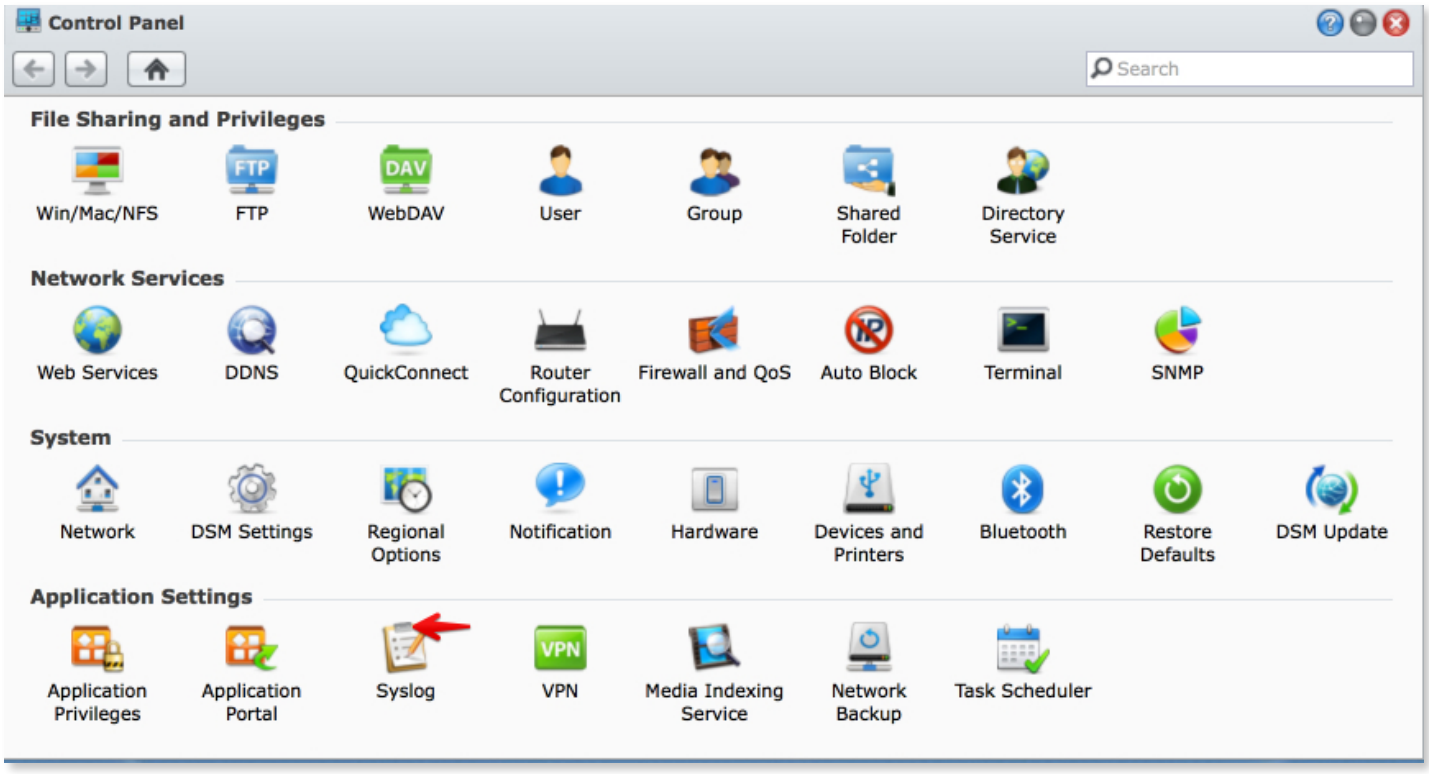
La solution de Synology fut ma première piste étudiée.

Temps de mise en place

Synology dispose de la solution qui est la plus rapide à mettre en place. Cette solution est pratiquement de type *ready out-of-the-box*. Son système d'exploitation est déjà installé, prêt à être configuré. Il est pourvu d'une interface graphique ressemblant beaucoup à l'interface Gnome 3 tout comme bon nombre de distributions Linux. Ses fonctionnalités serveur sont très faciles à configurer.



Synology DSM 4.3 GUI



Synology DSM 4.3 Control Pannel

Maintenance

La maintenance sur les NAS Synology est très simple. On peut configurer une sauvegarde sur un disque externe, ou encore un *High Availability Configuration* pour avoir une redondance sur deux serveurs Synology. Le système d'exploitation supporte la technologie S.M.A.R.T, décrivant en permanence l'état des disques durs. Le système d'exploitation effectue ses mises à jour automatiquement.

Fonctionnalités

Le serveur de stockage Synology intègre toutes les fonctionnalités voulues :

- Serveur VPN
- Serveur DNS / Serveur DHCP

- Serveur de fichiers FTP et serveur web HTTP
- Serveur multimédia iTunes
- Supporte la virtualisation Microsoft Hyper-V et VMware vSphere
- Serveur de sauvegarde Windows / Linux / Mac OS X

La plupart des fonctionnalités ne sont pas préinstallées. Elles sont plutôt offertes sous forme de package téléchargeable par l'utilisateur.

Configuration matérielle et spécifications

L'unité de stockage répondant à mes besoins est le Synology DiskStation DS1513+. L'unité est dotée des spécifications suivantes :

Composant/Spécification	Détail
Processeur	Intel Atom D2700 2.13 GHz • http://ark.intel.com/fr/products/59683/intel-atom-processor-d2700-1m-cache-2_13-ghz
Mémoire vive	• 2 GO DDR3 PC3-1066 Maximum 2x2 GB, total 4 GO
Nombre de baies	5 baies SATA 3.5"
Nombre de baies maximum avec expansion	15 baies maximum
Capacité de stockage maximale	20 TO
Ports externes	USB2 : 4 USB3 : 2 eSATA : 2
Connexions RJ-45	4x 1 Gbps LAN Intel WGI210AT
Support LACP Link Aggregation	✅
Wake On LAN (WoL)	✅
Ventilateurs	2 x 80mm
Alimentation	200w

Composant/Spécification	Détail
Consommation électrique	Environ 60w en charge Environ 30w au repos (*idle*)
Protocoles de communications supportés	CIFS AFP NFS FTP WebDAV CalDAV iSCSI Telnet SSH SNMP VPN (PPTP, OpenVPN)
Gestion	Auto DSM Upgrade Push Notification – MSN/Skype/Mobile Devices Email/SMS Notification Customized User Quota Customized Administrator/User Group Syslog Support DDNS Support IPv6 Support VLAN Support PPPoE Hotspot Resource Monitor Connection Manager UPS Management Scheduled Power On/Off Custom Management UI HTTP/HTTPS Ports
Gestion des disques	HDD Hibernation S.M.A.R.T Dynamic Bad Sector Mapping
Sécurité	FTP over SSL/TLS IP Auto-Block Firewall Encrypted Network Backup over Rsync HTTPS Connection
Taille maximale du système de fichier	108 TO
Nombre maximal de volumes internes	512 volumes
Nombre maximales de cibles iSCSI (*iSCSI targets*)	32
Nombre maximale de iSCSI LUNs	256
Possibilité de clones/Snapshots iSCSI	✅

Composant/Spécification	Détail
Types de RAID supportés	Synology Hybrid RAID Basic JBOD RAID 0 RAID 1 RAID 5 RAID 6 RAID 10
Migrations de RAID	Basic to RAID 1 Basic to RAID 5 RAID 1 to RAID 5 RAID 5 to RAID 6
SSD Cache en lecture	✅
SSD TRIM	✅
Virtualisation Microsoft Hyper-V	✅
Virtualisation VMware vSphere 5 avec VAAI	✅
Virtualisation Citrix	✅

Évolutivité

Au niveau du système d'exploitation, nous sommes limités par les packages de Synology. On notera toutefois qu'il existe déjà un bon nombre de package déjà disponibles. Pour une description de ces packages, visiter le lien ci-dessous.

- http://www.synology.com/dsm/dsm_app.php?lang=enu

Étrangement, le NAS Synology se distingue bien au niveau des possibilités futures malgré sa petite taille. Au niveau Ethernet, on ne peut pas faire vraiment mieux avec quatre interfaces Gigabit Ethernet déjà embarquées nativement sur le serveur. Synology marque des points sur l'expansion physique. On peut augmenter considérablement la capacité de stockage de ce NAS qui dispose de deux ports d'extension eSATA permettant de raccorder deux unités DX513, ajoutant un total de 10 baies supplémentaires. On peut cumuler jusqu'à 60 TO de capacité *RAW*

de stockage (sans RAID), ce qui est très impressionnant. Cependant, il m'a été impossible de déterminer si le processeur Intel Atom est suffisamment performant pour gérer le tout. Heureusement, Synology ajoute un contrôleur dans l'unité d'expansion DX513 qui enlève une partie de la charge de travail au processeur Atom. Il est tout de même intéressant de savoir qu'une telle chose est possible, et surtout que Synology gère l'expansion de batteries RAID *on-the-fly*, sans avoir à reformater la batterie de disque.

À l'exception de la capacité de stockage, on peut difficilement faire autre chose avec ce NAS. Le format du serveur est trop petit pour ajouter quelconque composant. Mais à regarder le DS1513+ sous ses angles matériels et logiciels, je ne vois personnellement pas ce que l'on pourrait rajouter de plus à long terme puisqu'il dispose déjà de bonnes plateformes pour chacun des deux aspects.

Rendement énergétique

Cette solution procure le meilleur rendement énergétique. Les composants utilisés dans ce NAS sont qualifiés comme étant à très faible consommation énergétique. Les processeurs Intel Atom sont les processeurs ayant la plus basse consommation de la lignée d'Intel. Avec une consommation électrique d'environ 60w pour l'unité en charge (*under load*), la consommation électrique annuelle ne sera pas très élevée et le coût d'opération par année pourra pratiquement passer inaperçu par rapport à la facture d'électricité actuelle.

Performances

Les performances de ce NAS sont principalement limitées à la bande passante des liens Gigabit Ethernet. En effet, un lien Gigabit Ethernet ne peut transférer, en théorie, plus de 125 MB/s brute. Avec l'ajout des entêtes de contrôle de chaque paquet et le trafic de contrôle engendré par les divers protocoles de communications, il ne reste qu'environ 110 MB/s de bande passante effective pour les données. Le NAS Synology n'a d'autres choix que d'accoter cette valeur.

On peut améliorer les performances avec des disques durs plus rapides de l'ordre de 7200 RPM. Ces disques durs offriront de meilleurs débits sur les tailles de fichiers plus petits, tout en offrant de bien meilleurs temps d'accès. Parmi les disques durs recommandés, on peut y voir les disques Seagate Constellations CS et ES.3 ainsi que les Western Digital SE et RE.

Ce NAS supporte, avec la dernière version du Synology DSM 4.3, le SSD caching. Cette technologie permet, dans le cas de Synology, de mettre en cache de lecture les fichiers

récemment et fréquemment utilisés par l'utilisateur. Fait à noter, c'est une cache de lecture et non pas d'écriture. Cela veut dire que le serveur ne peut, à l'heure d'écrire ces lignes, mettre en cache l'information que l'utilisateur écrit sur le NAS. Cela améliorerait grandement les performances en écriture de petits fichiers, car les temps d'accès des SSDs sont infiniment plus petits que ceux des disques durs conventionnels. Récemment dans le courant du mois de Janvier 2014, Synology a annoncé lors de la sortie de la version Bêta de DSM 5.0 que le SSD Caching en écriture sera désormais chose possible dans la prochaine version 5.0 du système. Le SSD caching est une fonctionnalité qui requiert deux SSDs mis en RAID 0. Cela limite donc grandement la capacité de stockage du NAS. Autre limitation : la cache SSD ne peut être effective que sur un seul volume au sein du NAS. La cache utilise un système FIFO - First-in-first-out - lorsque les SSDs sont remplis à pleine capacité.

Le NAS de Synology vient de base avec 2 GB de mémoire vive. La mémoire vive ne peut pas servir de cache. On peut mettre à niveau la mémoire pour avoir la possibilité de faire opérer plus de services en même temps sur le serveur. Un autre fait intéressant au niveau de la mémoire est que le système d'exploitation stock dans la mémoire vive la table d'allocation de la cache SSD. Cette table permet de savoir quels fichiers se trouvent dans les SSDs et à quel endroit en mémoire ils sont stockés. Cela fait en sorte que la taille de la cache SSD est directement proportionnelle avec la taille de la mémoire. De ce fait, des SSDs mis en RAID 0 peuvent rapidement créer un *overflow* dans la mémoire vive. Deux SSDs de 250 GO en RAID 0 font 500 GO de capacité de stockage. Mais avec seulement 4 GB de mémoire vive, seulement 164.53 GO peuvent être utilisés. On peut évidemment contourner le problème en ne mettant que des SSD de 100 GO qui totaliseront 200 GO en RAID 0, ce qui limitera les pertes. Toutefois, cela reste un très gros inconvénient de ce système, surtout dans des petits NAS où la capacité en mémoire vive est très limitée. Sur les gros NAS de la même compagnie qui opèrent sur des processeurs Intel Xeon avec, par exemple, 32 GO voir 64 GO de mémoire vive, le problème tend à être moins grave.

Au niveau des performances de transfert de données et de virtualisation, le NAS se débrouille très bien. Certes, le *SSD Caching* apporte un énorme plus lorsqu'on fait tourner des machines virtuelles à distance sur le NAS. Les machines virtuelles ont tendance à être beaucoup plus rapides et réagissent plus rapidement. Toutefois, on atteindra jamais les performances en virtualisation d'un système DAS (*Directly Attached Storage*), les liens Gigabit Ethernet n'étant pas assez rapides pour fournir la même rapidité que les bus internes d'un ordinateur. Cependant, le NAS exploite au maximum ses interfaces réseau.

Possibilités futures

Avec une plateforme plutôt limitée, le NAS Synology DS1513+ est restreint en terme de possibilités futures. Je pourrai toutefois faire actuellement tout ce que j'ai envie avec ce NAS au niveau virtualisation. Cependant, lorsque sera venu le temps d'entreprendre les certifications de VMware, il ne pourra plus suffire à la demande en performances. Même avec le système de *SSD Caching*, le NAS demeure relativement limité en terme de performances pour gérer des VHD (*Virtual Hard Disk*) dû au stockage à distance. Rappelons que le principal goulot d'étranglement demeure la connectivité Gigabit Ethernet. Certes, pour un petit parc de machines virtuelles par le service Microsoft Hyper-V dans le cadre des certifications du géant de Redmond, j'ai confiance qu'il pourra suffire à la demande en performance. Mais ce point reste à voir dans le cas de VMware. Il est toutefois compatible avec VMware vSphere. Cependant, pour étudier les certifications VMware, il me faudra un serveur de virtualisation ESXi, et ce NAS ne peut remplir cette tâche. Il pourra toutefois servir de stockage distant iSCSI pour ce serveur. Beaucoup utilisent ce NAS pour leur laboratoire personnel VMware à titre de stockage.

Sécurité

Au niveau de la sécurité du système d'exploitation, on est très loin de la sécurité des NAS de marque HP. Synology est très proactif dans les mises à jour. Le système d'exploitation tend à avoir un *kernel* Linux à jour. Toutefois, les versions des composants serveurs tels que PHP, MySQL ou Apache ne sont pas réellement tenus à jour fréquemment. De ce fait, n'importe quel serveur personnalisé aura une meilleure sécurité logicielle des applications serveur qu'un NAS Synology (mais également de n'importe quelle autre marque de NAS avec système d'exploitation intégré et mis à jour par le fabricant). Cependant, il faut également se dire que lorsqu'on loue un serveur web ou un espace d'hébergement, les versions utilisées de LAMP (Linux, Apache, MySQL/Maria DB, PHP/Pearl) sont rarement les dernières versions disponibles. De ce fait, le niveau de sécurité est similaire.

Il ne faut pas oublier le niveau de sécurité des données. Le NAS de Synology utilise à l'interne le système de fichier EXT4. Ce système n'inclue donc pas toutes les technologies de sécurité du ZFS, par exemple la copie sur écriture ou encore le *silent data corruption repair*. À long terme, il est donc impossible de prédire la qualité des données enregistrées, comme tout autre système de fichier commun.

Conclusion

Le NAS Synology DS1513+ est une excellente machine. On notera ses performances admirables, ses quatre liens Gigabit Ethernet, le support de VLANs et du protocole LACP 802.3ad. Il est également possible de lui ajouter jusqu'à deux unités d'expansion. Il est extrêmement facile à configurer et supporte les technologies de virtualisation. Toutefois, sa configuration logicielle est limitée, de même que son expansion matérielle quasi inexistante. Cela reste toutefois un alternative de choix et de haute qualité offrant des performances très raisonnables pour le coût engendré.

Étude de cas #2 : Serveur personnalisé

L'autre possibilité étudiée est celle d'un serveur personnalisé, assemblé sur mesure. Cette solution est, après celle d'un NAS préassemblé, la plus utilisée et la plus populaire au sein des administrateurs. Elle permet en effet d'administrer entièrement le système, tant au niveau matériel que logiciel.

Temps de mise en place

Le temps de mise en place de ce genre de serveur, qui serait alors un serveur de virtualisation, est très long. Contrairement à Synology où l'ensemble est livré dans un état prêt à opérer, un serveur de ce type n'est même pas assemblé. On doit en premier lieu choisir les composants. Par la suite, un hyperviseur doit y être installé, en l'occurrence, VMware ESXi avec vSphere 5.5. Une fois cet hyperviseur installé, un système d'exploitation destiné au stockage (avec préférablement le système de fichier ZFS puisque la configuration matérielle le permet) doit être installé dans une machine virtuelle. Le stockage se trouve alors disponible après avoir paramétré ce système virtualisé. Toutefois, il reste encore à l'optimisé, puisque le ZFS demande une certaine optimisation de ses niveaux de cache. Somme toute, cette solution serait un système au final très performant, mais très complexe et très long à mettre en place.

Maintenance

Si le système est long à mettre en place, la maintenance serait quant à elle moins fastidieuse. En fait, la maintenance repose essentiellement sur les mises à jour du système

d'exploitation de stockage utilisé (FreeNAS, OpenIndianna, Solaris, Illumos, etc). Et même si des mises à jour sont disponibles, la complexité du système fait en sorte que l'administrateur sera plutôt réticent à l'égard de faire une mise à niveau du système logiciel. Toutefois, ces systèmes sont très souvent issus d'une communauté OpenSource (à l'exception de Solaris qui appartient à la multinationale Oracle) et donc très bien documentés par la communauté. Le support est toutefois aussi assuré par celle-ci, et non pas par un vendeur officiel comme Synology. Ces systèmes sont souvent bâtis sur des kernel ayant fait leurs preuves en terme de stabilité.

Fonctionnalités

C'est ici où ce type de serveur se distingue. En effet, puisqu'il s'agit d'un serveur de virtualisation, on peut le personnalisé comme bon nous semble. On peut assurer divers services par l'intermédiaire de différentes machines virtuelles, toutes connectées au réseau local virtuel et physique. Dans un tel système, le seul facteur limitant s'avère la configuration matérielle, qui peut elle limiter le nombre de machines virtuelles pouvant être opérées en même temps. On peut très bien faire de la ségrégation de services par machine virtuelle opérant sur différents systèmes d'exploitation. Un serveur de ce type n'est pas limité par des packages précis comme le cas de Synology puisqu'il peut avoir tous les packages disponibles sous les branches Debian et Red Hat. Ce n'est pas, à proprement parler, un serveur dédié qu'au stockage de fichiers, comme a tendance à faire Synology.

Configuration matérielle et spécifications

C'est ici où le tout commence à être dispendieux. En effet, comme mentionné lors de la couverture du ZFS, ce système de fichier prend énormément de ressources physiques. Si on veut que notre serveur soit efficace pour autre chose que la machine virtuelle jouant le rôle de NAS/SAN, il faut miser haut sur le processeur et la mémoire vive. Cette dernière doit être également de type ECC. De ce fait, la configuration du serveur de stockage pourrait ressembler à ceci :

Composant	Spécification
Processeur	Intel Xeon E5-1650 V2 (Ivy Bridge - E)
Carte-mère	X9SRH-7F
Mémoire vive (RAM)	Kingston ValueRAM ECC 1600 MHz 32 GB 4x8 GB KVR16R11S4K4/32I Registred

Composant	Spécification
Interfaces LAN	Intel i350 Dual-Port on-board Gigabit Ethernet Realtek RTL8201N on-board Gigabit Ethernet (dedicated IPMI)
Boitier	SuperMicro
Alimentation	Built-in PSU
Stockage système d'exploitation	Intel Business SSD DC S3700 100 GB
Stockage de fichiers	Western Digital SE 4 TB (x3 minimum)
Stockage de cache	Intel Business SSD DC S3700 100 GB
Stockage des machines virtuelles	Samsung 840 Evo 1 TB / Crucial M500 960 GB
HBA (Host Bus Adaptor)	On-Board LSI 2308 Possibility to add another LSI HBA
Vidéo	On-board Matrox G200eW

Une telle configuration est bien évidemment très performante, mais porte également son prix. Toutefois, du tel matériel permet bien des possibilités, incluant même la certification VMware vSphere, Microsoft et Red Hat.

Évolutivité

La configuration d'un serveur de ce type permet la plus grande évolutivité possible, puisque le format du serveur est standardisé, nous ne sommes pas limités en terme d'expansion par fentes PCI/PCI-Express et SATA. Le seul facteur limitant est le nombre de fentes PCI/PCI-Express de la carte mère pour les expansions de ce type ainsi que le nombre d'emplacements pour disques durs du boitier (puisque l'ont peut ajouter des *SAS expanders*).

Rendement énergétique

Au niveau consommation énergétique, on est très loin de celle d'un NAS Synology. De tels composants ne peuvent forcément que consommer plus d'énergie. La plateforme elle-même

consomme environ 150w, consommation à laquelle il faut ajouter celle des disques durs. Toutefois, au niveau rendement, qui prend en compte les performances de la machine, elle est exemplaire. Ivy-Bridge E (Xeon) a actuellement le meilleur rendement énergétique de sa catégorie. Toutefois, puisqu'un tel serveur consomme davantage qu'un NAS, le coût opérationnel sera plus élevé et ne passera pas aussi inaperçu que la solution Synology.

Performances

Au même titre que le NAS Synology, les performances de ce système pour la partie serveur de fichier sont principalement limitées à la bande passante des liens Gigabit Ethernet, soit environ 110 MB/s.

Toutefois, le système de fichier ZFS permet la cache en écriture, diminuant énormément les temps d'accès en écriture sur le NAS, chose que Synology ne fait pas pour le moment. Le ZFS gère également le cache en lecture. On peut facilement atteindre plusieurs dizaines de milliers d'IOPS avec ce genre de serveur. On peut donc dire qu'un tel serveur aurait de meilleures performances globales quant au système de fichier, mais que celui-ci est également plus lourd à gérer et requiert par le fait même plus de ressources matérielles.

Cependant, ce système n'a pas de vocation entière à un serveur de fichier. Il est serveur de virtualisation, donc il peut aussi opérer n'importe quel autre système d'exploitation. Une telle configuration matérielle pour ce genre d'usage s'avère parfaite et offrira les meilleures performances possible.

Possibilités futures

Il est certain que ce type de serveur offre les meilleures possibilités futures puisqu'il est malléable autant du côté logiciel que matériel. Il me permettrait de faire les certifications VMware tout en me servant de serveur de fichier.

Sécurité

La sécurité des données est sans aucun doute la meilleure qui soit puisque celles-ci sont stockées sur un système de fichier ZFS. C'est sans l'ombre d'un doute le meilleur système possible pour assurer la sécurité des données à long terme.

Pour ce qui est des autres services, le tout est configuré dans des machines virtuelles Linux. L'administrateur est donc responsable d'assurer la sécurité des services et applications Web. L'administration au niveau sécurité est facile puisqu'on peut mettre à jour les différentes applications telles qu'Apache, MySQL et PHP ainsi que le système d'exploitation aussitôt qu'une mise à jour est disponible. La sécurité est donc beaucoup plus proactive dans ce cas-ci qu'avec Synology où l'on dépend des efforts de la firme en terme de mise à jour du système.

Conclusion de l'analyse

Voici un tableau récapitulatif de l'analyse sous chaque point.

Point étudié	Synology	Serveur de virtualisation	Gagnant
Temps de mise en place	★★★★★	★	Synology
Maintenance	★★★★	★★★	Synology
Fonctionnalités	★★★★	★★★★★	Serveur
Configuration matérielle	★★	★★★★★	Serveur
Évolutivité	★★	★★★★★	Serveur
Rendement énergétique	★★★★★	★★	Synology
Performances (système de fichier)	★★★★	★★★★★	Serveur
Possibilités futures	★★★★	★★★★★	Serveur
Sécurité	★★★	★★★★★	Serveur
Coût total	★★★★★	★	Synology

Temps de mise en place

Le fait que Synology soit prêt *out-of-the-box* le fait drastiquement gagné dans ce comparatif. Il est très facilement mis en place et en très peu de temps, contrairement à un serveur où chaque point devra être étudié, installé puis configuré manuellement.

Maintenance

La maintenance est plus simple chez Synology puisque les mises à jour se font automatiquement. Maintenir les services d'un serveur peut s'avérer être un job en soit. Synology a réussi à rendre faciles la maintenance et l'administration des fonctionnalités de son unité.

Fonctionnalités

Bien que Synology ait un grand nombre de fonctionnalités, le gagnant est sans aucun doute le serveur puisqu'on peut configurer n'importe quel service sur n'importe quel système d'exploitation virtualisé.

Configuration matérielle

Le serveur en sort évidemment gagnant vu sa configuration matérielle beaucoup plus performante que celle du NAS Synology.

Évolutivité

Le serveur sort gagnant sur ce point puisqu'on peut aisément et rapidement changer sa configuration matérielle, alors que le NAS Synology utilise un format propriétaire dont la seule spécification modifiable est la mémoire vive, et ce dans une plage limitée à 4 GO de mémoire. Aucune carte d'extension n'est possible.

Rendement énergétique

Le prix du rendement énergétique au point de vue consommation revient sans l'ombre d'un doute à Synology avec sa consommation qui passe pratiquement inaperçue sur la facture d'électricité, alors que le serveur peut être très énergivore. Toutefois, au point de vue du ratio performance/consommation, le serveur en sort gagnant avec ses composants haut de gamme et très performant à ce niveau. Cependant, nous ne pouvons pas réellement comparer les deux solutions sur ce point, puisque celles-ci ont deux vocations différentes.

Performances (système de fichier)

Les performances à ce niveau sont très semblables. Toutefois, le serveur de fichier sur le système ZFS a une fonctionnalité de cache supplémentaire : la cache en écriture. On peut deviner qu'il aura des performances globalement meilleures. Cependant, on reste limité par le lien Gigabit Ethernet, donc le gain en performance se situe dans une plage plutôt limitée.

Possibilités futures

Le serveur en sort légèrement gagnant à ce niveau puisqu'il me permettrait d'effectuer des certifications VMware, Microsoft, Red Hat et autres.

Sécurité

Puisque chaque service et application web est administré manuellement par l'administrateur, ce dernier a plein contrôle sur la sécurité et les mises à jour critiques de ces composants. Le serveur en sort donc gagnant, mais au prix d'une plus grande maintenance de la part de l'administrateur.

Prix

Nous arrivons donc au seul point non discuté lors de l'analyse ; le prix.

Ces équipements coûtent cher dans les deux cas. Il faut plutôt le voir comme étant un investissement, puisqu'ils serviront dans les deux cas à accroître mes connaissances.

Dans le cas du NAS Synology, l'unité elle-même, sans les disques durs, se détaille au prix de 800 $ CND au moment d'écrire ces lignes. À cela, j'ajouterais trois disques durs pour le RAID 5. Ces disques durs sont des Western Digital SE 4 TO qui se détaillent au prix de 275 $ l'unité. Si on veut ajouter les SSD de *caching*, je dois ajouter au NAS deux SSD Samsung ou Crucial qui se détaillent entre 100 et 140 $ chaque. Le total maximal de ce système reviendrait alors à 2000 $ CND.

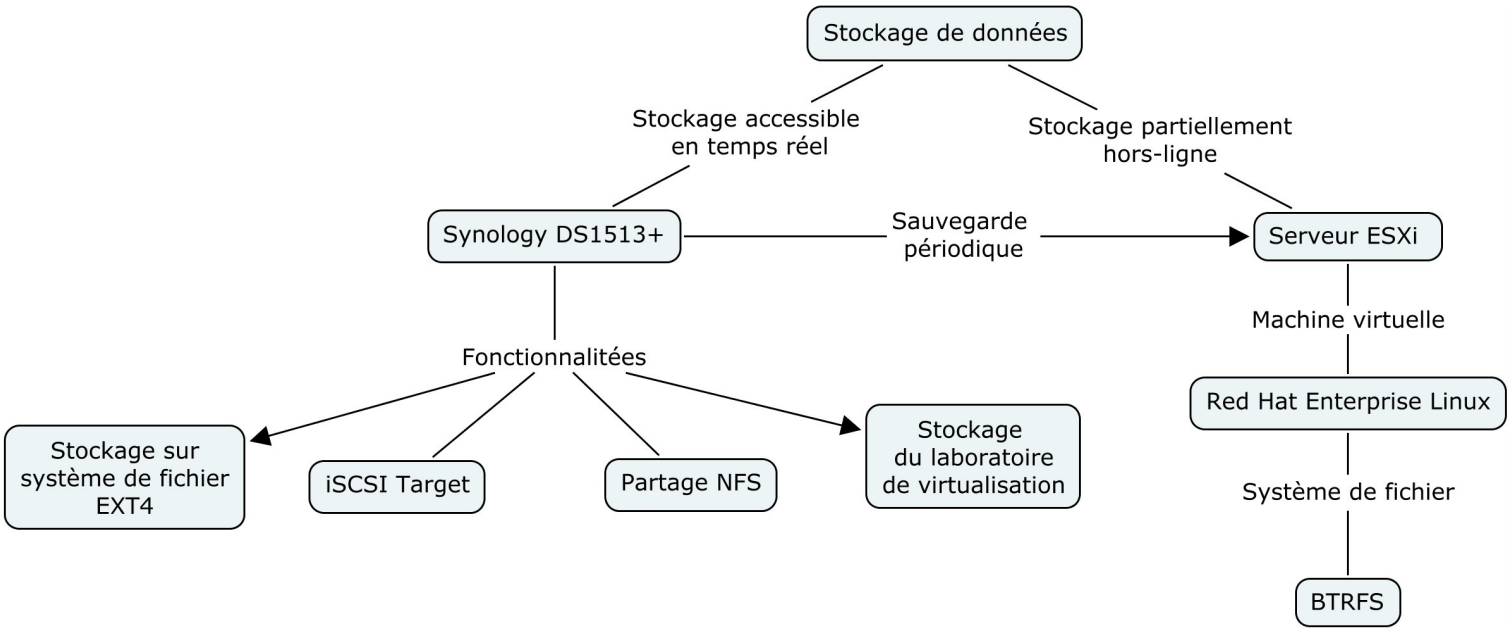
Dans le cas du serveur, une simple estimation sommaire du coût d'une telle configuration matérielle reviendrait à plus de 2500 $ CND sans les disques durs. Le coût total serait donc de plus de 3000 $ CND, soit un peu moins que le double du prix du Synology.

Conclusion de l'achat

La solution achetée fut le Synology DiskStation DS1513+.

Ce choix peut paraître contradictoire avec l'issue du comparatif décrit dans la conclusion, le comparatif étant en faveur du serveur personnalisé. Le principal point ayant penché dans la balance fut le coût total de la mise en place de l'installation. L'achat du NAS Synology, est revenu exactement à 2000 $ CND. À cela, je devais rajouter un 250 $ CND supplémentaire pour me procurer un UPS (*uninterruptible power supply*) 1500VAC pour préserver l'intégrité des données et du serveur en cas de panne de courant.

L'option envisagée à long terme est la suivante :



Un serveur ESXi serait acheté dans le futur lorsque j'envisagerai de compléter des certifications. Ce serveur, dont une machine virtuelle servira de machine de sauvegarde du pool Synology, fonctionnera sur le système de fichier BTRFS sous Red Hat Enterprise Linux, lorsque ce dernier intègrera une version stable du système de fichier. Ce serveur pourra être doté de matériel moyen de gamme puisque le stockage ne servira que de sauvegarde et ne sera pas activement sollicité. Je pourrai alors me construire un *ESXi host* à un coût plus abordable, en prenant par exemple une carte mère et un processeur *server grade* de milieu de gamme. Le NAS Synology pourra toujours servir de stockage distant à ce serveur si le besoin est présent.

Le choix de Synology a également été fait à cause de la simplicité de cette solution. Certes, elle est moins performante et moins sécuritaire, mais elle est clef en main et dispose de moins de temps de maintenance, tout en ayant de très bonnes performances et un choix de fonctionnalités varié.

Durant mon étude de cas, j'ai par ailleurs testé dans des machines virtuelles les distributions FreeNAS, Illumos Napp-It et Solaris. Bien que les deux premières simplifient la gestion du système de fichier ZFS, j'avais l'impression de faire tourner un système peu sécuritaire. Dans le premier cas, l'interface est très belle et fonctionnelle, mais n'intègre pas toutes les fonctionnalités de Synology. Bien que la distribution soit stable, plusieurs utilisateurs ont des problèmes avec FreeNAS lorsque virtualisé. La distribution pour entreprise TrueNAS semble moins sujette à ces problèmes récurrents et dispose d'une plateforme beaucoup plus complète et solide. Napp-It, contrairement à FreeNAS, semblait complexifier la gestion du ZFS, tout en portant ce même sentiment de "*home-made system*" dont la fiabilité sur le papier est certes très bonne, mais douteuse à l'usage. Solaris possède de loin les meilleures performances et le meilleur sentiment à l'usage, mais revient à faire un serveur à partir de zéro et n'intègre pas d'emblée une fonction primordiale à mon usage : Apple Time Machine (il faut savoir que Time Machine sur Solaris *peut être* possible, mais avec quelques heures de bricolage autour du système d'exploitation). En résumé, seul FreeNAS constituait pour moi une véritable solution si le ZFS avait été employé. Toutefois, quelques tests de performance en machine virtuelle avec un *pass-through* direct à un disque dur physique ne m'ont guère enchanté et j'ai compris que beaucoup de RAM ECC ainsi que des SSD de caching étaient absolument de mise dans ce genre de système pour avoir des performances décentes tout en gardant les fonctionnalités et bienfaits du ZFS.

Devant ce constat, il ne restait que le serveur ESXi tournant sur Debian ou Red Hat Enterprise Linux (RHEL) qui puisse être envisageable si le ZFS devait être employé. *ZFS on Linux*, bien que distribué sous forme de version stable, représente un véritable défi à instaurer et à maintenir. Une simple mise à jour du kernel Linux peut nécessiter la reconfiguration des pools de stockage (sans perte de données, heureusement). Finalement, j'ai conclu que le serveur ESXi tournerait probablement sur RHEL (ou une variante de RHEL) ou FreeNAS sous le système de

fichier EXT4 ou UTF dans le cas de FreeNAS. Les performances auraient sensiblement été les mêmes que dans le cas d'une solution Synology, sans toutefois être une solution clef en main, ni offrir aucune sécurité supplémentaire à celle de Synology.

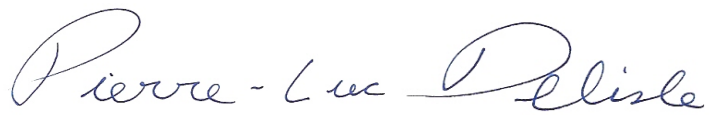
Pour conclure, Synology représente bien des avantages et le produit livré clef en main et très stable a fait pencher la balance. Le choix fut très difficile, probablement un des plus difficiles que j'ai dû faire puisque toutes les possibilités, même le DAS, étaient parfaitement envisageables et réalisables. Mais la simplicité et le coût de Synology l'ont emporté au terme d'une longue réflexion.

Conclusion

Ceci conclut mon essai sur les technologies de stockage mises en réseau.

Pour plus de renseignements concernant les points traités dans cet essai, n'hésitez pas à consulter la bibliographie disponible à la page suivante.

Merci beaucoup de votre lecture.



Pierre-Luc Delisle

Bibliographie

Computer World, Lucas Mearian, Western Digital's HAMR tech could increase disk capacity five-fold. [en ligne].
http://www.computerworld.com/s/article/9244052/Western_Digital_s_HAMR_tech_could_increase_disk_capacity_five_fold?source=cwfb

Le Journal du Net. Le stockage informatique à l'heure de la virtualisation. [en ligne].
http://www.journaldunet.com/solutions/systemes-reseaux/stockage/

Intel Communities, Only 1Gbps on a 2Gbps Team Dynamic Link (802.3ad) w/ PRO1000 PT & PM Port Adapters. [en ligne].
https://communities.intel.com/thread/19600

IEEE802.org. IEEE 802.3ad Link Aggregation - What it is, what it is not. [en ligne].
http://www.ieee802.org/3/hssg/public/apr07/frazier_01_0407.pdf

Jason Nash's Blog, Jason Nash, Synology Category. [en ligne].
http://jasonnash.com/category/synology-2/

HardForum Community, various threads. [en ligne].
http://hardforum.com/showthread.php?p=1040387390&posted=1#post1040387390
http://hardforum.com/showthread.php?t=1791679
http://hardforum.com/showthread.php?t=1764125&highlight=
http://hardforum.com/showthread.php?t=1783020&highlight=link+aggregation

Synology Corporation, DiskStation Manager Packages. [en ligne].
http://www.synology.com/dsm/dsm_app.php?lang=enu

Wikipedia, SAN - Storage Area Network. [en ligne].
http://en.wikipedia.org/wiki/Storage_area_network

Wikipedia, Link Aggregation. [en ligne].
http://en.wikipedia.org/wiki/Link_Aggregation

Wikipedia, Transmission Control Protocol. [en ligne].
http://en.wikipedia.org/wiki/Transmission_Control_Protocol

Wikipedia, B-Tree. [en ligne].
http://fr.wikipedia.org/wiki/Arbre_B

Wikipedia, mdadm. [en ligne].
http://en.wikipedia.org/wiki/Mdadm

Wikipedia, Storage Virtualization. [en ligne].
http://en.wikipedia.org/wiki/Storage_virtualization

Wikipedia, RAID. [en ligne].
http://en.wikipedia.org/wiki/RAID

Wikipedia, Storage Hypervisor. [en ligne].
http://en.wikipedia.org/wiki/Storage_hypervisor

PacketLife.net, EtherChannel considerations. [en ligne].
http://packetlife.net/blog/2010/jan/18/etherchannel-considerations/

Thomas-Krenn.com, Link Aggregation and LACP basics. [en ligne].
http://www.thomas-krenn.com/en/wiki/Link_Aggregation_and_LACP_basics

Synology Corporation, Synology Forum, Using Link Aggregation on the Synology DiskStation. [en ligne].
http://forum.synology.com/wiki/index.php/Using_Link_Aggregation_on_the_Synology_DiskStation

Synology Corporation, Synology Forum, Configure Link Aggregation with Cisco Switch IEEE 802.3ad [en ligne].
http://forum.synology.com/enu/viewtopic.php?f=145&t=72571

Synology Corporation, Synology Forum, How to use iSCSI Targets on VMware ESXi with Multipath I/O. [en ligne].
http://forum.synology.com/wiki/index.php/
How_to_use_iSCSI_Targets_on_VMware_ESXi_with_Multipath_I/O

Cisco Corporation, EtherChannel Between a Cisco Catalyst Switch That Runs Cisco IOS and a Workstation or Server Configuration Example. [en ligne].
http://www.cisco.com/en/US/tech/tk389/tk213/
technologies_configuration_example09186a008089a821.shtml

Cisco Corporation, Configuring EtherChannel and Link-State Tracking, Configuring Layer 2 EtherChannel. [en ligne].
http://www.cisco.com/en/US/docs/switches/lan/catalyst2960/software/release/12.2_53_se/
configuration/guide/swethchl.html#wp1275918

Cisco Corporation, Catalyst 4500 Series Switch Cisco IOS Command Reference, 12.2(20)EWA, Port-channel load-balance. [en ligne].
http://www.cisco.com/en/US/docs/switches/lan/catalyst4500/12.2/20ewa/command/reference/int_sess.html#wp1977735

Cisco Corporation, Understanding EtherChannel Load Balancing and Redundancy on Catalyst Switches. [en ligne].
http://www.cisco.com/en/US/tech/tk389/tk213/technologies_tech_note09186a0080094714.shtml

Blog Open-E.com, Janusz Bak, Bonding versus MPIO Explained. [en ligne].
http://blog.open-e.com/bonding-versus-mpio-explained/

Microsoft Technet, Understanding MPIO Features and Components. [en ligne].
http://technet.microsoft.com/en-us/library/ee619734(v=ws.10).aspx

Microsoft Technet, MPIO Policies. [en ligne].
http://technet.microsoft.com/en-us/library/dd851699.aspx

Windows IT Pro, John Howie, Microsoft Multipath I/O for iSCSI. [en ligne].
http://windowsitpro.com/storage/microsoft-multipath-io-iscsi

IT Infrastructure Solutions, Gurav Anad, What is MPIO and Best Practices of MPIO configuration. [en ligne].
http://itinfras.blogspot.ca/2010/05/what-is-mpio-and-best-practices-of-mpio.html

Dell AppAssure. Repository Options: Direct Attached Storage, Storage Area Network or Network Attached Storage? [en ligne].
http://www.appassure.com/support/KB/repository-options-direct-attached-storage-storage-area-network-or-network-attached-storage/

BTRFS Wiki. [en ligne].
https://btrfs.wiki.kernel.org/index.php/Main_Page

The Solidq Journal, Joe Chang, I/O Queue Depth Strategy for Peak Performance. [en ligne].
http://www.solidq.com/sqj/JournalDocuments/2011_January_Issue/sqj%20007%20pag.%2024-31.pdf

NetApp.com, Ryan Hardin, Windows Multipathing Options with Data ONTAP : Fibre Channel and iSCSI, Technical Report. [en ligne].
http://www.netapp.com/us/system/pdf-reader.aspx?m=tr-3441.pdf&cc=us

Altaro blog for Hyper-V and Windows Administrators, Eric Siron, Teaming and MPIO for Storage in Hyper-V 2012. [en ligne].
http://www.altaro.com/hyper-v/teaming-and-mpio-for-storage-in-hyper-v-2012/

VMware | Blogs, Eric Horschman, Our position on hypervisor footprints, patching, vulnerabilities and whatever else Microsoft wants to throw into a blog post. [en ligne].
http://blogs.vmware.com/virtualreality/2009/08/our-position-on-hypervisor-footprints-patching-vulnerabilities-and-whatever-else-microsoft-wants-to-throw-into-a-blog-post.html

InformationIsBeautiful.net, Codebases. [en ligne].
http://www.informationisbeautiful.net/visualizations/million-lines-of-code/

arstechnica.net, Jim Salter, Bitrot and atomic COWs: Inside “next-gen” filesystems. [en ligne]
http://arstechnica.com/information-technology/2014/01/bitrot-and-atomic-cows-inside-next-gen-filesystems/#image-2