# Système en haute disponibilité sous Linux
# SAN (Storage Area Network), Cluster Suite et Global File System

Fabien Muller
Institut de Physique et de Chimie des Matériaux de Strasbourg
23 Rue du Loess 67034 Strasbourg Cedex
Fabien.Muller@ipcms.u-strasbg.fr

Hubert Hollender
Institut de Physique et de Chimie des Matériaux de Strasbourg
23 Rue du Loess 67034 Strasbourg Cedex
Hubert.Hollender@ipcms.u-strasbg.fr

## Résumé

*Cet article décrit la mise en œuvre d'un réseau de stockage en haute disponibilité à l'Institut de Physique et de Chimie des Matériaux de Strasbourg.*

*La solution mise en place est basée sur la technologie SAN (Storage Area Network), avec 2 serveurs reliés via des interfaces internes à une baie RAID Fibre Channel, l'ensemble étant totalement redondant. L'utilisation optimale des liens Fibre Channel entre la baie et les serveurs (charge IO, disponibilité) est assurée par le logiciel « Powerpath ». Chaque nœud du cluster accède simultanément au même espace de stockage grâce à Global File System (GFS), un système de fichiers en grappe. La gestion de la haute disponibilité des services est prise en charge par le logiciel « Cluster Suite ». Le cluster fonctionne en mode actif/actif. L'accès aux services se fait à partir des informations stockées dans un annuaire LDAP répliqué sur les deux serveurs. Pour les postes clients sous Linux, PAM (Pluggable Authentication Modules) et autofs sont utilisés. Pour les postes clients sous Windows un couplage LDAP-SAMBA a été mis en place.*

*Le bilan d'exploitation de cette solution a permis de démontrer sa grande robustesse. L'utilisation conjointe de Cluster Suite avec GFS permet la construction de clusters en environnement actif/actif hautement fiabilisés.*

## Mots clefs

SAN, Cluster Suite, GFS, Powerpath, RAID, LDAP, haute disponibilité

## 1 Introduction

Cet article décrit la mise en œuvre d'un réseau de stockage en haute disponibilité à l'Institut de Physique et de Chimie des Matériaux de Strasbourg. L'IPCMS regroupe des physiciens et des chimistes dont l'objectif est de concevoir et d'élaborer de nouveaux matériaux. L'institut dispose d'un système d'information performant composé actuellement de plus de 500 machines (moyens destinés au calcul scientifique, serveurs centraux, postes de travail, équipements réseaux).

Pour faire face à une utilisation sans cesse croissante des nouvelles technologies de l'information, se traduisant par la génération de volumes de données de plus en plus importants et sensibles, la direction de l'institut a très vite compris l'importance de la mise en place d'une infrastructure de stockage centralisée et sécurisée. Ainsi, dès 1998, une première architecture propriétaire de haute disponibilité a été mise en œuvre.

La solution décrite dans cet article constitue l'évolution de cette architecture initiale en améliorant ses points faibles, à l'époque son système de stockage basé sur une baie RAID logicielle et en prenant en compte l'augmentation constante de la masse de données à gérer et leur criticité croissante. Elle est basée sur la technologie SAN (Storage Area Network). La haute disponibilité est assurée via le logiciel « Red Hat Cluster Suite » et le système de fichiers en grappe « Global File System ».

## 2 Architecture mise en œuvre

L'architecture matérielle est constituée d'une baie SAN [1] Fibre Channel interconnectée via des liens Fibre Channel rebondants de 2 Gb/s à deux serveurs fonctionnant en cluster actif/actif. La partie sauvegarde est assurée par une librairie LTO-3 attachée en SCSI à un des deux serveurs du cluster et le logiciel Time Navigator.

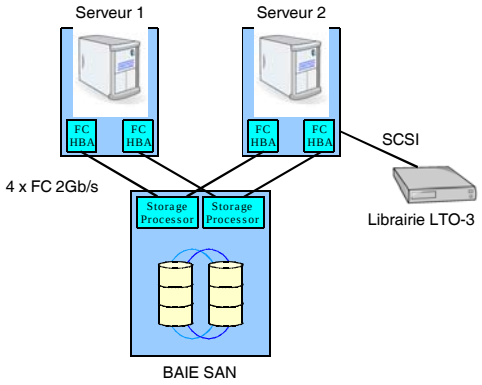


Figure 1 – *Architecture matérielle*

## 2.1 La baie SAN

La baie assure une continuité de service maximale grâce à l'utilisation massive de composants redondants. Elle dispose d'une architecture matérielle où tous les éléments sont, au moins, doublés. En plus des alimentations et des ventilateurs redondants, elle est équipée de deux processeurs de stockage (SP) qui sont connectés via quatre boucles FC-AL indépendantes aux serveurs. Chaque processeur peut desservir chaque disque de la baie via l'une ou l'autre de ces boucles. Deux liens fibre channel internes à 2 Gb/s, les CMI (communications messaging interface) relient les deux contrôleurs. Chaque disque étant muni d'une double connexion FC-AL, un incident sur une des deux boucles est sans conséquence sur l'accès aux données. Globalement aucun composant matériel ne peut empêcher le fonctionnement du système.

L'utilisation de batteries de secours intégrées et d'un mécanisme sécurisé des caches, garantissent l'intégrité des données contre les pannes électroniques et contre les coupures d'alimentation électrique. Toutes les informations entrant dans le cache d'un SP en écriture sont copiées directement sur la partie miroir du cache en écriture de l'autre SP. La taille mémoire en lecture et en écriture des caches peut être paramétrée pour optimiser les performances en entrée et sortie. Un espace spécial réservé sur les cinq premiers disques, nommé « vault » est configuré en RAID3. Inaccessible aux utilisateurs, il permet d'avoir un espace persistant pour les données résidant dans le cache en écriture, en cas d'événement grave pouvant compromettre l'intégrité des données. Lorsqu'un problème survient comme une coupure générale d'électricité un dump du cache est effectué (copie d'une image binaire incluant toutes les pages du cache et les pointeurs).

Dans sa configuration actuelle, la baie est équipée de 15 disques Fibre Channel hot plug de 300 Go, dont 1 est utilisé en hot spare. Deux RAID group de 7 disques en RAID 5 et 9 LUN (5 et 4 par RAID Group et SP) ont été configurés. La capacité utile de stockage est de 3,6 To.

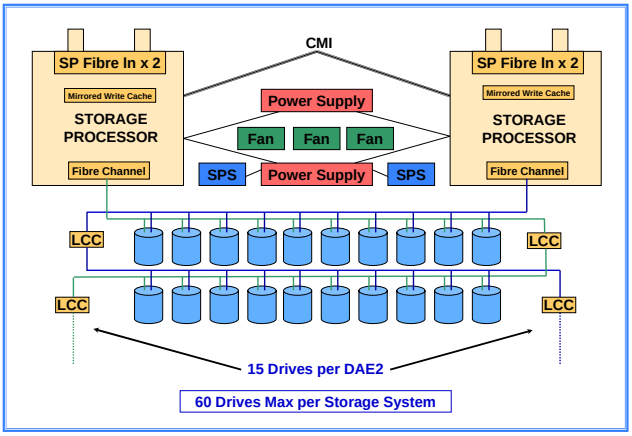


Figure 2 – *Architecture du SAN*

Pour pouvoir administrer de façon optimale la baie, une solution logicielle nommée « Navisphere » est fournie. Il s'agit d'une suite de management du constructeur, basée sur une interface web (applet Java) permettant la gestion et la configuration de la baie, le suivi des incidents et le contrôle d'applications optionnelles.

## 2.2 Les serveurs

La configuration des deux serveurs à base de biprocesseurs Xéon à 3,2 GHz est identique. Ils intègrent 2 Go de mémoire, un contrôleur RAID SCSI ultra 320 permettant la gestion de 6 disques Hot plug, 2 interfaces Gigabit Ethernet et 2 cartes Qlogic Fibre channel. Les alimentations électriques et les ventilateurs sont hot plug et redondants. La fonction Memory Spare Bank permet de disposer d'une barrette de mémoire spare. Au niveau système Linux « Red Hat Advanced Server » est utilisé.

## 2.3 Powerpath

Le logiciel Powerpath installé sur chaque serveur, améliore les performances et la disponibilité des informations. Powerpath offre des fonctions de gestion dynamique de la charge et de basculement des trajets sur incident entre les serveurs et la baie. En cas de perte d'un chemin d'accès d'un serveur vers la baie, Powerpath redirige automatiquement les demandes I/O vers le ou les canaux restants. Il effectue aussi l'équilibrage de charge sur les différents canaux, ce qui permet d'amener un gain de performance significatif. Powerpath réside au-dessus des drivers HBA et permet la gestion de 2 à 32 canaux.

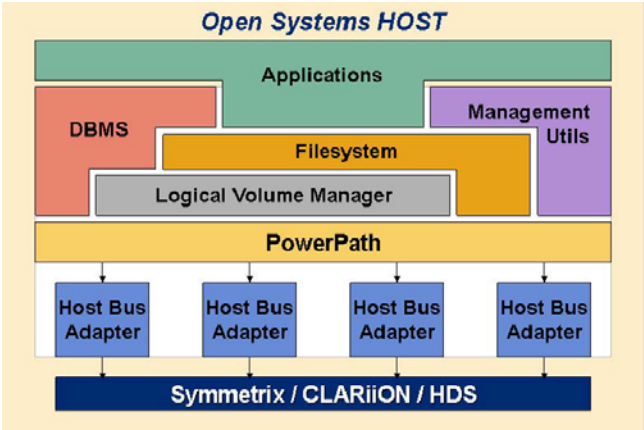


Figure 3 – *Powerpath*

# 3 Gestion de la haute disponibilité

La gestion de la haute disponibilité des services est assurée par une technologie de clustering. Un cluster est un agrégat de machines permettant la réalisation d'un travail coopératif. Les clusters peuvent être utilisés pour augmenter la puissance de traitement (scalability) dans le cas des clusters de calcul ou pour augmenter la disponibilité (availability) et ainsi minimiser les

inconvénients liés aux pannes par la redondance des machines entre elles.

Dans ce cas, le cluster sera composé de 3 sous-systèmes logiques, l'accès réseau constituant le point de passage entre les serveurs du cluster et les postes de travail, le support du système de fichier (en général une baie de disques partagés ou éventuellement les disques locaux des machines), le cœur de calcul composé de n couples mémoire-CPU. Un cluster haute-disponibilité doit obligatoirement pouvoir supporter un arrêt et un redémarrage brutal d'un service sans que cela ne soit perceptible par les utilisateurs. Les services doivent aussi être monitorables pour connaître à tout instant leurs états fonctionnels.

Il existe différents modes de cluster. Le plus simple à mettre en œuvre est le mode actif-passif. Dans ce cas un serveur est dédié à la reprise de services. Dans le mode actif/actif tous les nœuds du cluster tournent des services. Une seule instance est active à un instant donné et en cas de défaillance le service est repris par une des autres machines. Dans ce mode il y a des risques de corruptions de données et il faut prévoir des mécanismes pour éviter cela. Enfin, il existe les clusters à répartition de charge, appelés aussi fermes de serveurs. Dans ce cas le même service est réparti sur plusieurs machines, l'accès pour les postes clients se faisant via un serveur frontal lui même sécurisé en mode actif/passif.

La solution mise en œuvre à l'institut fonctionne en cluster actif/actif et s'appuie sur le logiciel « Red Hat Cluster Suite » et le système de fichiers en grappe « Global File System ».

## 3.1 Global File System (GFS)

Global File System (GFS) [3] est un système de fichiers en grappe (clustered file system), accessible simultanément de plusieurs serveurs d'un cluster. Implanté à l'origine sous Irix, GFS a été porté par Sistina Software, après que SGI ait mis les sources de son système de fichiers dans le domaine public. GFS supporte la journalisation et la récupération des données perdues consécutivement à la défaillance d'un des membres du cluster.

Les nœuds d'un cluster GFS partagent physiquement le même espace de stockage par le biais de la fibre optique ou de périphériques SCSI partagés. Le système de fichiers semble être local sur chaque nœud et GFS synchronise l'accès aux fichiers sur le cluster. GFS est complètement symétrique ce qui signifie que tous les nœuds sont équivalents et qu'il n'y a pas un serveur susceptible d'être un entonnoir ou un point de panne. GFS utilise un cache en lecture écriture tout en conservant la sémantique complète du système de fichiers Unix.

La synchronisation pour l'accès aux données entre les machines du cluster est assurée par un mécanisme de verrouillage distribué, DLM (Distributed Lock Manager) intégré dans la partie Core Services du produit Cluster Suite. La gestion des volumes logiques est assurée par la couche logicielle CLVM (Cluster Logical Volume Manager).

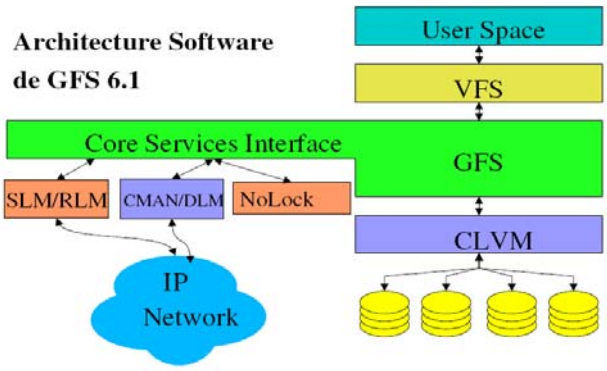


Figure 4 – *Architecture de GFS*

## 3.2 Cluster Suite

Red-Hat Cluster Suite [2] est un logiciel de clustering intégrant deux sous-ensembles logiciels. La première partie, Cluster Manager, permet la gestion de la haute disponibilité. La deuxième partie permet l'équilibrage de la charge IP sur les nœuds d'une ferme de serveurs. Cluster Suite est un sous-ensemble de GFS.

Le gestionnaire de grappe de cluster suite (cluster manager) permet de gérer au maximum 16 nœuds, garantit l'intégrité complète des données via l'utilisation de barrière I/O (fencing) et offre un service failover permettant la détection des défaillances matérielles et logicielles, les arrêts/remises en route automatique, et le contrôle des applications. Cluster manager s'appuie sur un concept de services. Chaque service à sécuriser est identifié par un nom, une adresse IP de service qui basculera avec le service de serveur en serveur, d'éventuels points de montage et un ensemble de scripts pour le démarrage et l'arrêt.

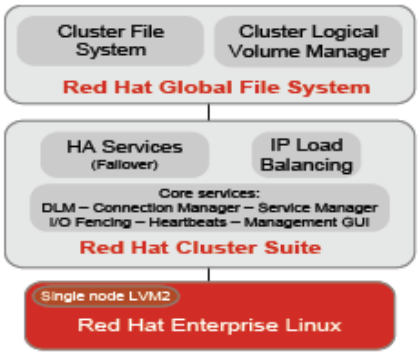


Figure 5 – *Architecture de Cluster Suite*

Pour garantir l'intégrité des données quoiqu'il arrive et éviter des écritures concurrentes de données, il faut en cas de panne sortir un nœud du cluster en l'arrêtant physiquement. Il est donc nécessaire de disposer d'une architecture hardware permettant de prendre le contrôle d'un serveur à distance et de l'éteindre via le réseau. Cette fonctionnalité est fournie dans cluster suite via des agents de fencing.


Figure 6 – Le fencing

# 4 Services sécurisés

Le cluster fonctionnant en mode actif/actif les services ont été répartis de façon équitable sur les deux serveurs. Neuf volumes logiques GFS ont été créés. Sept sont utilisés comme espace de stockage des utilisateurs, pour les services SAMBA et NFS (3 et 4 volumes GFS par serveur SAMBA). Pour les impressions des postes sous windows un troisième serveur SAMBA a été activé. Tous ces services SAMBA/NFS sont sécurisés via Cluster Suite. Le service CUPS quant à lui a été démarré sur les deux serveurs, les clients accédant par défaut à celui tournant la partie SAMBA.

Un autre volume GFS sert au stockage des boîtes aux lettres de la messagerie. L'architecture du courrier électronique de l'institut a été construite avec le logiciel Postfix. Un serveur installé dans la zone démilitarisée (DMZ) fait office de relais de messagerie vers le cluster, d'antivirus et d'anti-spam via Mailscanner, Sophos et Spamassassin. Pour la gestion des boîtes aux lettres des utilisateurs le service Cyrus-Imap a été clusterisé.

Enfin, un dernier volume GFS héberge les fichiers de configurations et d'erreurs des services sécurisés (samba, cups, cyrus, cluster suite) et évite ainsi des synchronisations en cas de modification.

# 5 L'environnement LDAP

La refonte de l'architecture du réseau de stockage de l'institut a été aussi l'occasion de migrer l'authentification et la gestion des droits des utilisateurs, l'accès aux ressources partagées, la gestion des adresses de messagerie électronique, des aliases et des listes de diffusion vers un annuaire LDAP. Pour les postes clients sous Linux, PAM (Pluggable Authentication Modules) et aufofs sont utilisés. Pour les postes clients sous Windows, un couplage LDAP-SAMBA a été mis en place. Pour garantir une disponibilité totale du service, l'annuaire a été répliqué sur les deux serveurs du cluster. Pour administrer cet annuaire une application web dynamique a été développée en php, « phpilam ».

# 6 Conclusion

Le bilan d'exploitation de cette solution qui est maintenant opérationnelle depuis dix huit mois a permis de démontrer sa grande robustesse. L'utilisation massive au niveau matériel de composants redondants associée à une technologie de clustering assure une continuité de service maximale. Le couple Cluster Suite et GFS permet la construction de clusters en environnement actif/actif hautement fiabilisés.

Dans sa version actuelle, la solution mise en place s'appuie au niveau système sur la version Linux Red-Hat Advanced Server et pour Cluster Suite et GFS sur leurs versions commerciales. Cependant, la même architecture logicielle peut être construite en utilisant une distribution libre et les versions open source des logiciels utilisés. Avec la baisse actuelle des prix des baies SAN, cela permet d'envisager la mise en exploitation d'infrastructures de haute disponibilité à des coûts bien plus raisonnables que par le passé.

Au niveau des évolutions futures envisagées, l'acquisition de deux commutateurs Fiber Channel externes permettra l'interface de la baie SAN avec nos différentes machines de calcul. Cette extension donnera aussi la possibilité d'augmenter le nombre de serveurs du cluster si cela s'avérait nécessaire. L'étude des technologies de virtualisation est aussi en cours. En effet, l'intégration de Xen dans Red-Hat 5 permet la construction de solutions de virtualisation complètes bien adaptées aux systèmes en grappes.

Enfin, la sécurisation générale de tous les services, même ceux qui ne sont pas clusterisés pour l'instant comme le DHCP ou la base de l'annuaire LDAP, via le volume GFS dédié au système, est en cours d'analyse.

# Bibliographie

[1] Bernard Debord, Retour d'expérience SAN multi site: de problèmes en solutions, Dans *Actes du congrès JRES2005*, pages 251-259, Marseille, Décembre 2005

[2] *Configuring and Managing a Red Hat Cluster*, Presse Red Hat, Mai 2007.

[3] *Red Hat Global File System*, Presse Red Hat, Janvier 2007.