



# Systeme en haute disponibilité

Fabien Muller

System Area Network et ClusterSuite




# Plan de l'exposé

- Contexte
- Approches
- Technologie SAN
- Technologie de clustering
- Solution mise en œuvre
- Bilan

# Problématique de départ

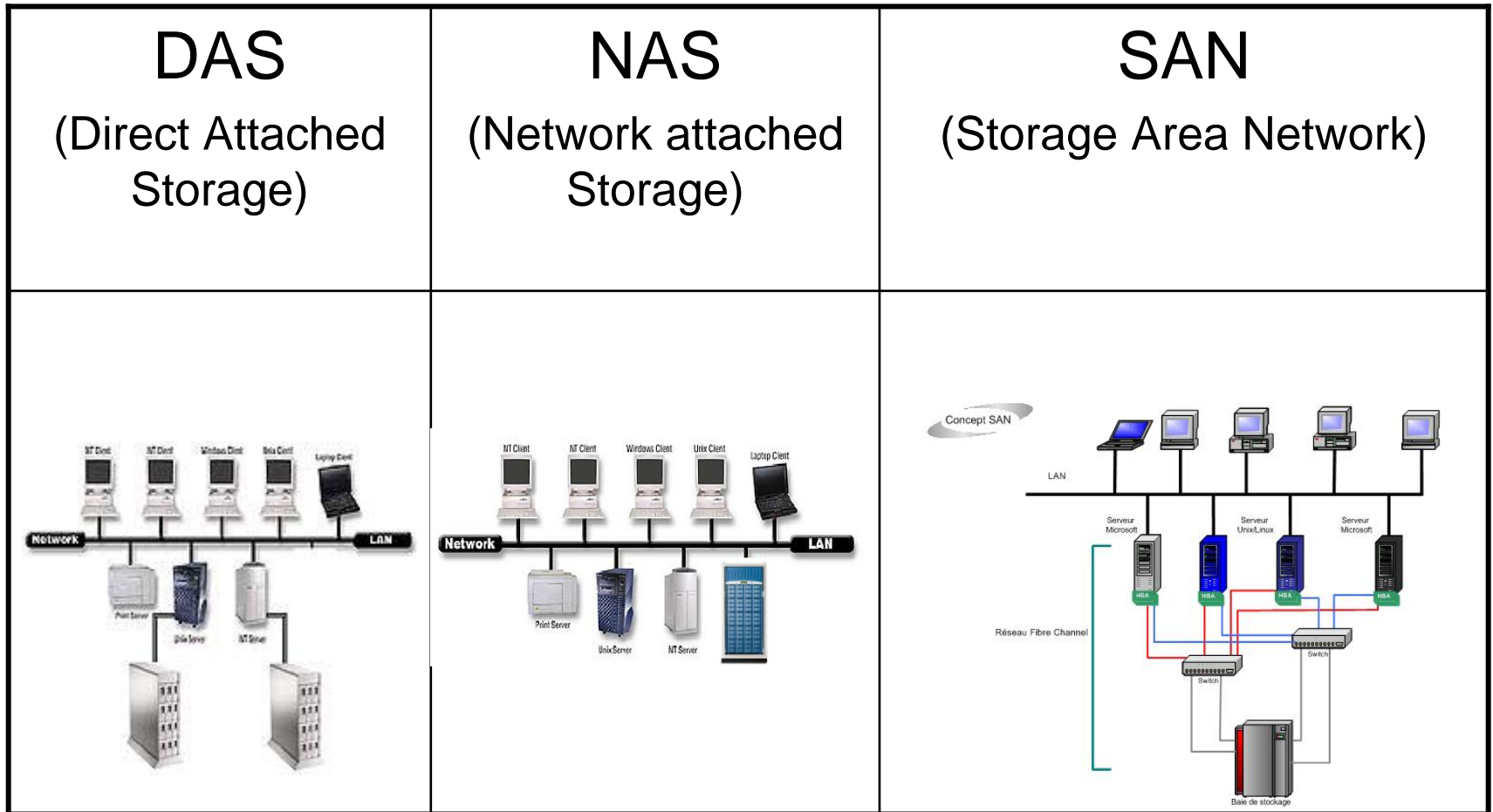
- Renouvellement de l'ancien système
- Volume de données
  - de plus en plus important
  - de plus en plus sensibles
- Temps d'indisponibilité de plus en plus réduit
- Réseau dédié pour le stockage
- Cluster à haute disponibilité



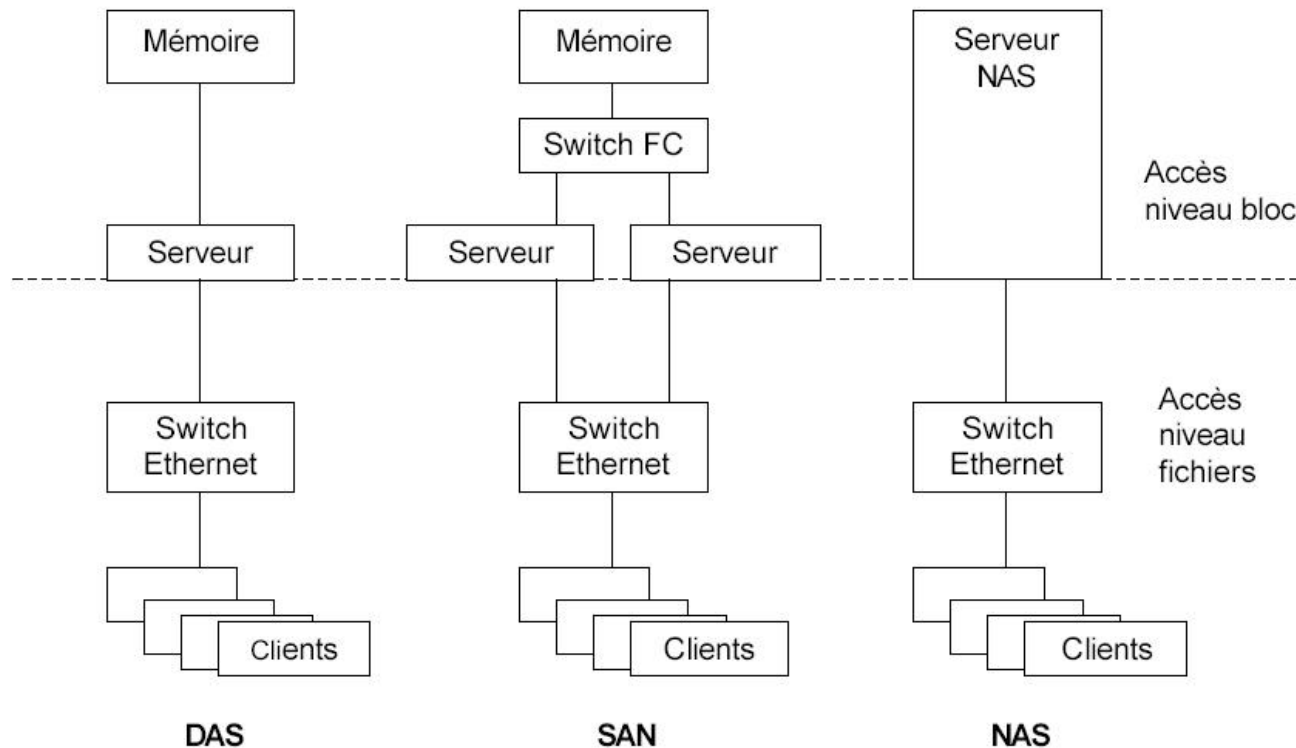
# Plan de l'exposé

- Contexte
- **Approches**
- Technologie SAN
- Technologie de clustering
- Solution mise en œuvre
- Bilan

# Architectures



# Architectures



- Accès au niveau fichier: CIFS dans le monde Windows (ex SMB), NFS dans le monde Unix
- Accès niveau block: « raw ». Les serveurs accèdent aux périphériques par une interface de bas niveau, via le protocole SCSI, quelque soit l'OS (hors mainframes).

# Plan de l'exposé

- Contexte
- Approches
- **Technologie SAN**
- Technologie de clustering
- Solution mise en œuvre
- Bilan

# SAN - Présentation

Historiquement le protocole le plus adapté pour le stockage est SCSI mais :

- Interface parallèle → débits limités (320 Mbps voire 640)
- Nombre de périphériques limité à 16. Longueur de câble de quelques mètres

Alors l'ANSI travaille sur un successeur à SCSI :

- SCSI-3 → serial SCSI
- Plus de limitation sur le nombre de périphériques (16 millions)
- Possibilité de connecter plusieurs serveurs sur un même périphérique
- Distances associées aux nouveaux support physiques: FC jusqu'à 10 Kms

## Qu'est qu'un SAN ?

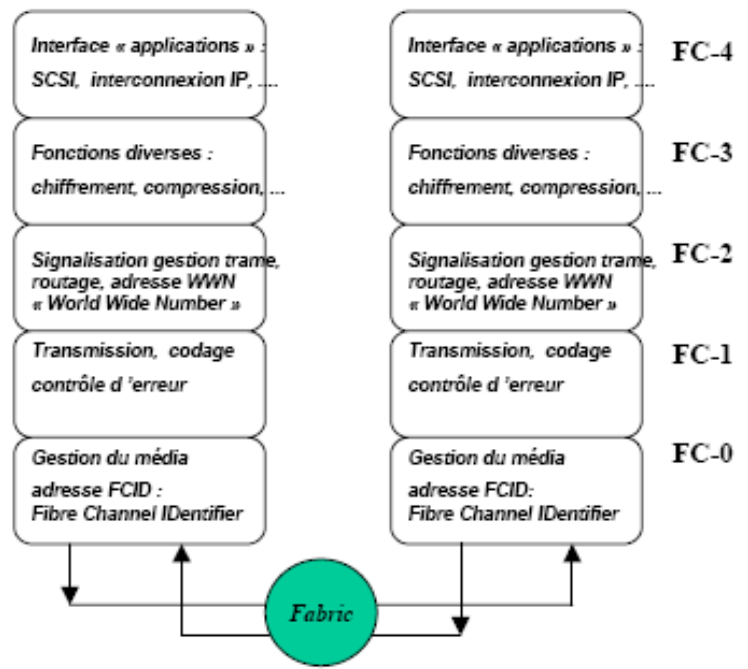
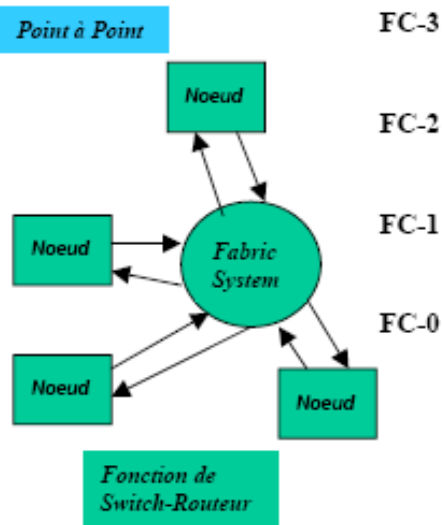
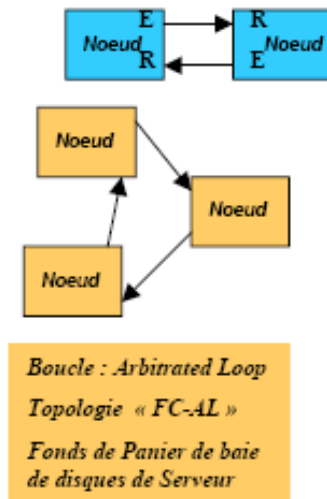
- Sur un réseau dédié ou non
- Sur Fibre optique ou sur Ethernet
- Protocole SCSI (FCP ou iSCSI)

**Définition : 2 périphériques ou plus communiquant par le protocole Serial SCSI (Fibre Channel ou iSCSI)**



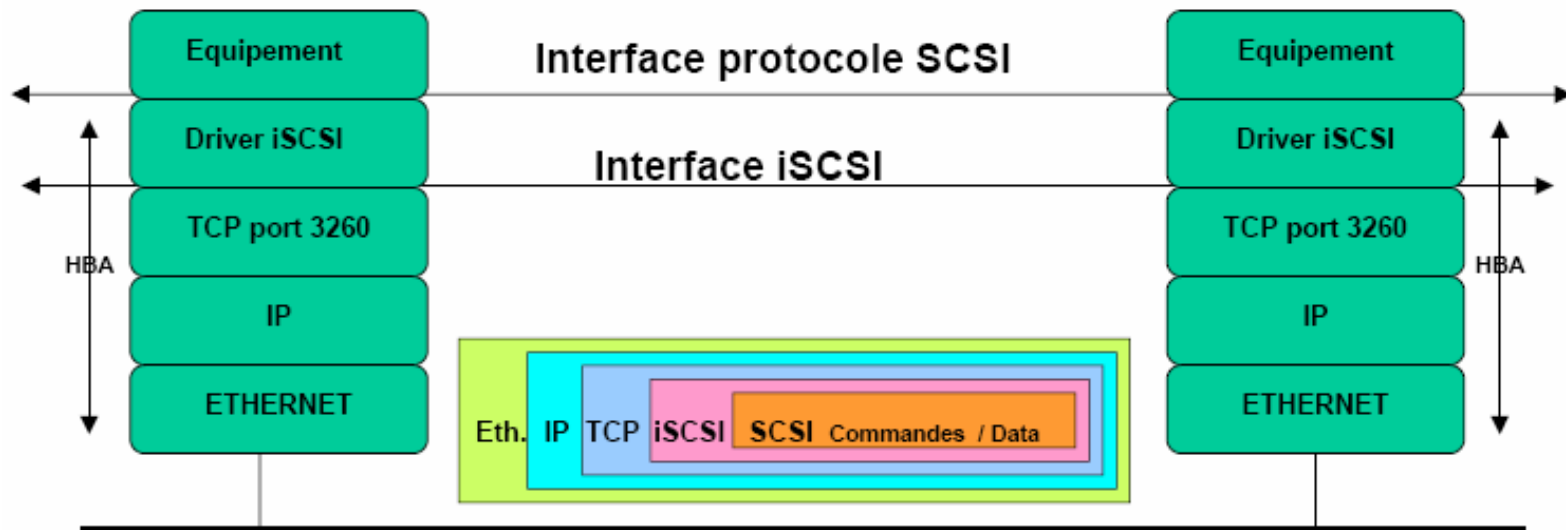
# SAN – Fiber Channel

- Conçu pour transporter le protocole SCSI (version SCSI-3)
  - transport à longue distance (> 10 km sur F.O.)
  - Sérialisation Full-Duplex du protocole de BUS // Half-Duplex
  - Interface à 2 ports : émission et réception
  - différentes topologies possibles
  - protocoles en 5 couches



# SAN - iSCSI

- **iSCSI** : *Internet Small Computer System Interface*
  - Initié par CISCO en 2000, normalisé par l'IETF en 2003
  - Encapsulation des trames SCSI dans TCP
  - Coupleur HBA (Host Bus Adaptateur) sortie Ethernet RJ45 directe
  - Permet de mettre en place un réseau de stockage « bon marché »
  - Pas de réseau dédié (mais performances « moindres » pour les applications critiques)



# Plan de l'exposé

- Contexte
- Approches
- Technologie SAN
- **Technologie de clustering**
- Solution mise en œuvre
- Bilan

# Présentation des Clusters

- Cluster = agrégat de machines dans un but de travail coopératif.
- Cluster : pour 2 fonctionnalités
  - Augmentation de la puissance de traitement (scalability) : on veut que la puissance de traitement suive de manière linéaire le nombre de machines du cluster.
  - Augmentation de la disponibilité (availability): on veut minimiser les inconvénients liées aux pannes par la redondance des machines entre elles.

# Cluster Haute-Disponibilité

- Le cluster est composé de 3 sous-systèmes logiques :
  - l'accès réseau : c'est le point de passage entre les machines du cluster et les machines clientes
  - le support du système de fichier :
    - baie disque partagées (SCSI / Fiber Channel)
  - Le coeur de calcul : n couples mémoire-CPU.
  
- Obligatoirement :
  - le service doit pouvoir supporter :
    - un arrêt brutal.
    - un redémarrage brutal.

# Cluster Actif-Passif

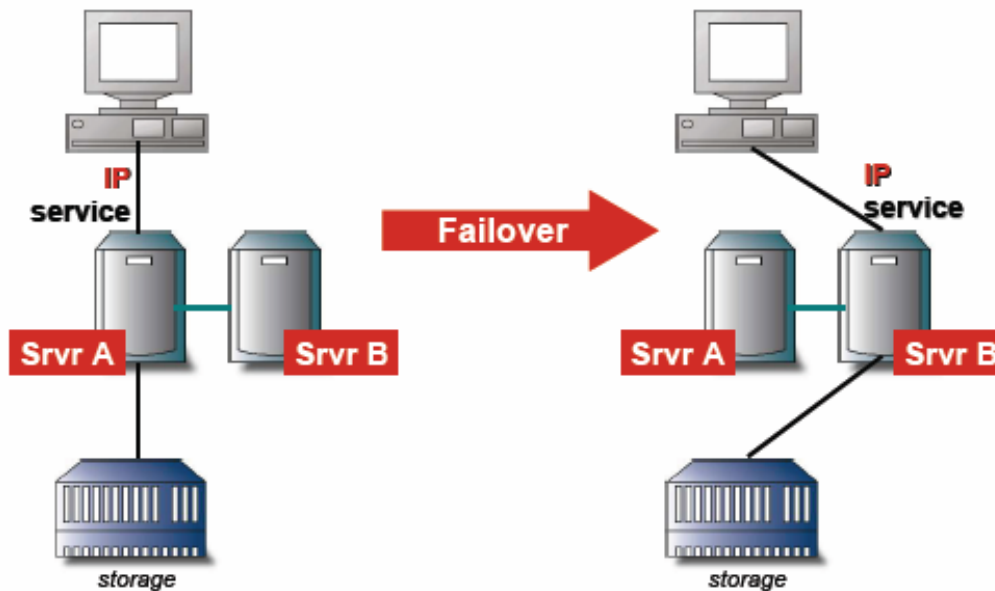
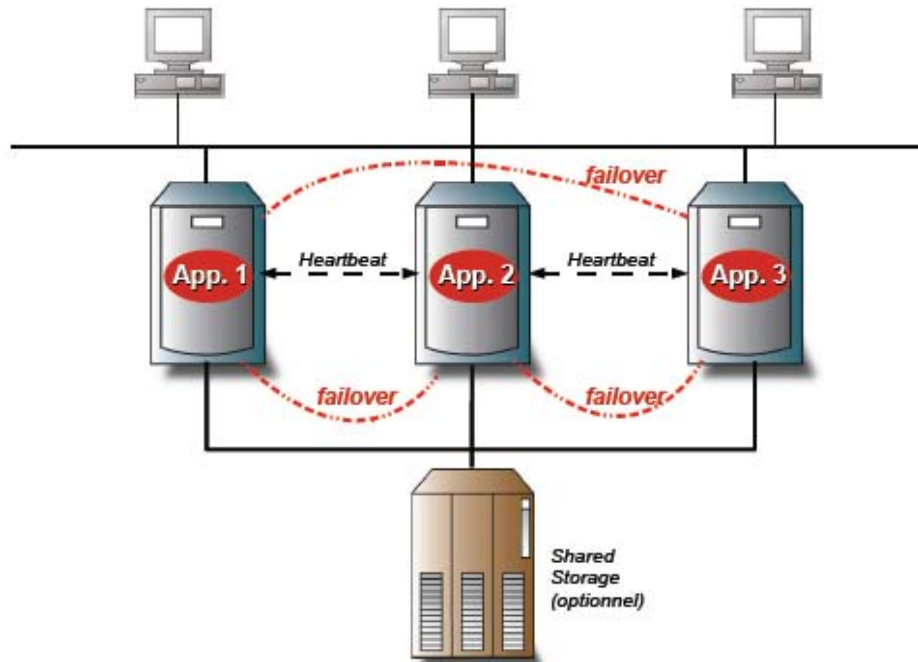


Illustration Séquence Illustration: failover

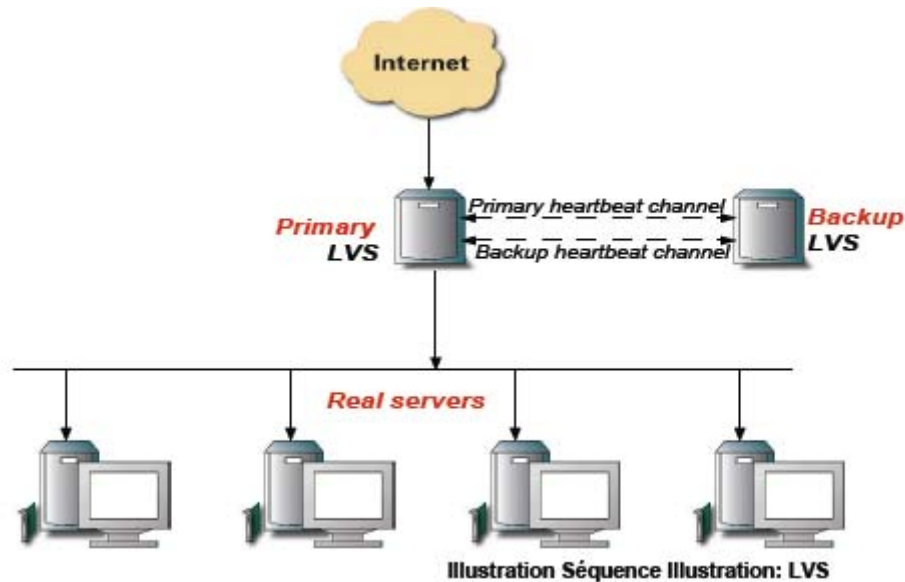
- ❑ Déploiement simplifiée, niveaux de performances garanti
- ❑ Serveur dédié à la reprise de services

# Cluster Actif-Actif



- ❑ tous les nœuds du cluster tournent des services
- ❑ une seule instance active du service
- ❑ risque de corruption des données (R/W simultanées)

# Cluster à répartition de charges



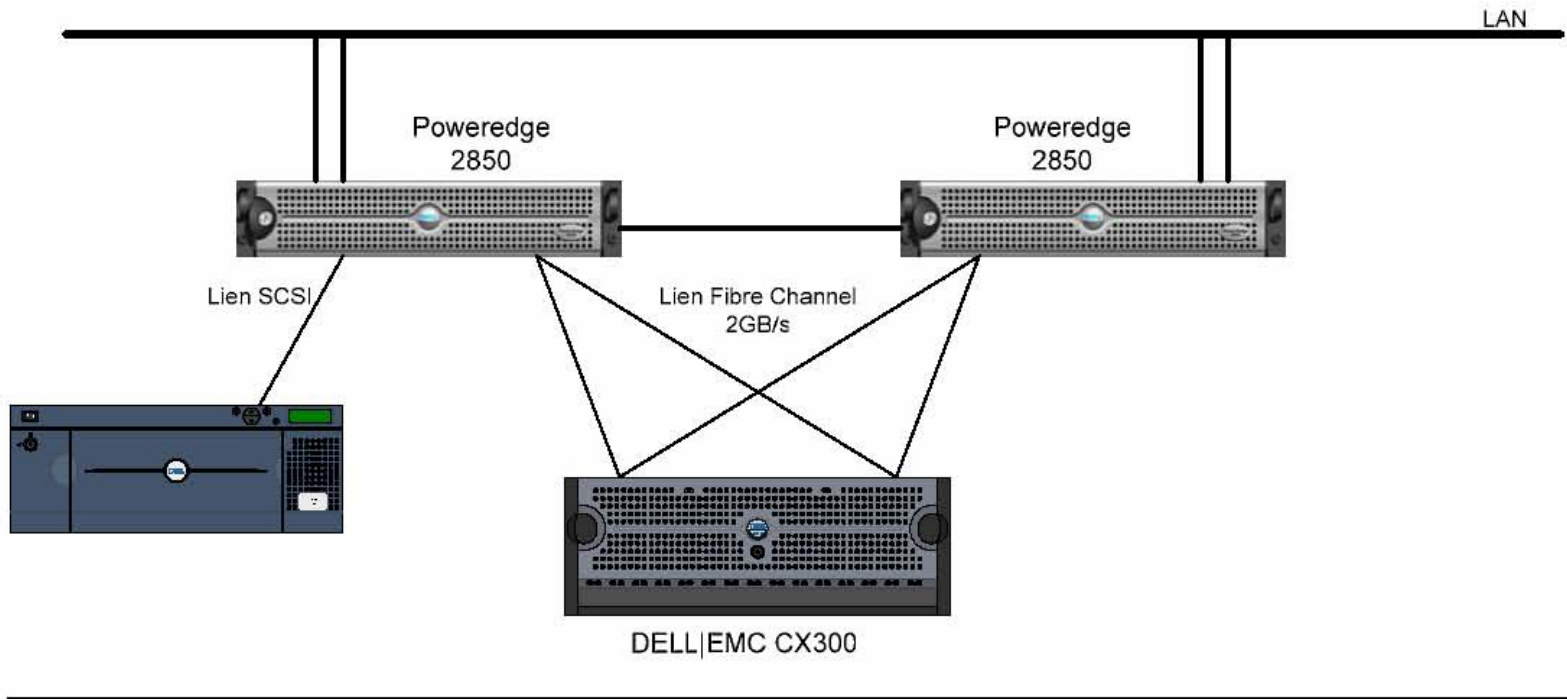
- ❑ Ferme de serveurs
- ❑ Répartition d'un même service sur plusieurs machines
- ❑ Pour le monde extérieur : un serveur unique



# Plan de l'exposé

- Contexte
- Approches
- Technologie SAN
- Technologie de clustering
- **Solution mise en œuvre**
- Bilan

# Solution mise en œuvre



# Baie disque : CX300

DAE2



DAE2



DAE2



DPE2



SPS

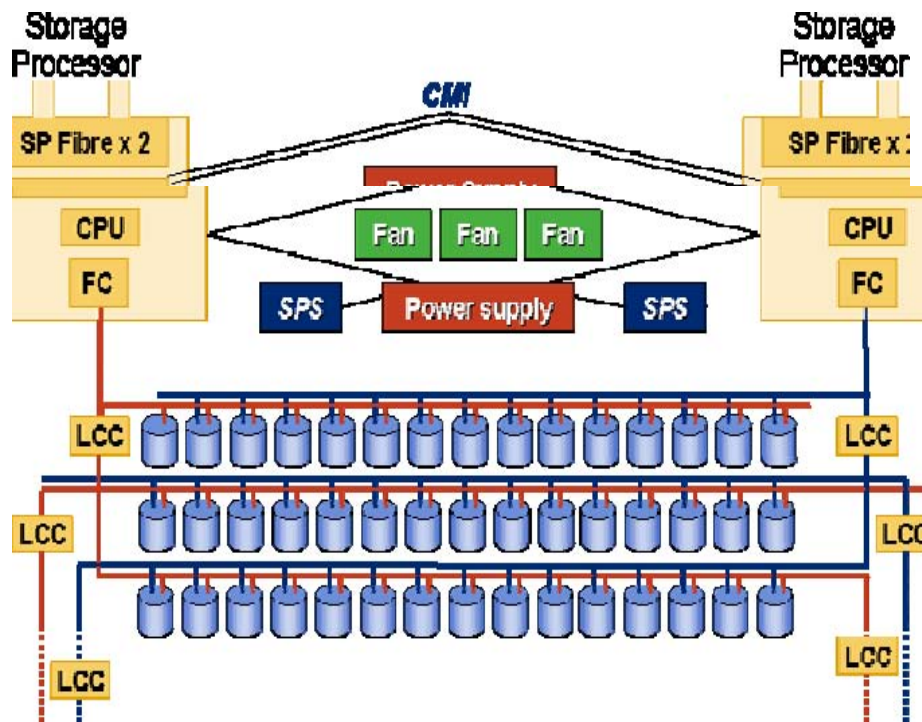


Configuration  
Minimum

- 15 disques par DAE2 et DPE2
- 60 disques max (18 To)  
Fiber Channel, ATA ou SATA
- 64 serveurs

Configuration  
Maximum

# Baie disque : CX300



- 2 processeurs à 800 MHz (SP)
- 2 Go de mémoire cache
- 4 connexions serveurs 2 Gb
- 4 bus disque interne 2 Gb
- 50 000 I/Os

# Baie disque : CX300

## ■ Cache :

- I/O miroré entre les Storage processeurs
- Paramétrage en lecture et écriture
- Paramétrage de la taille de page (2 à 16 Ko)

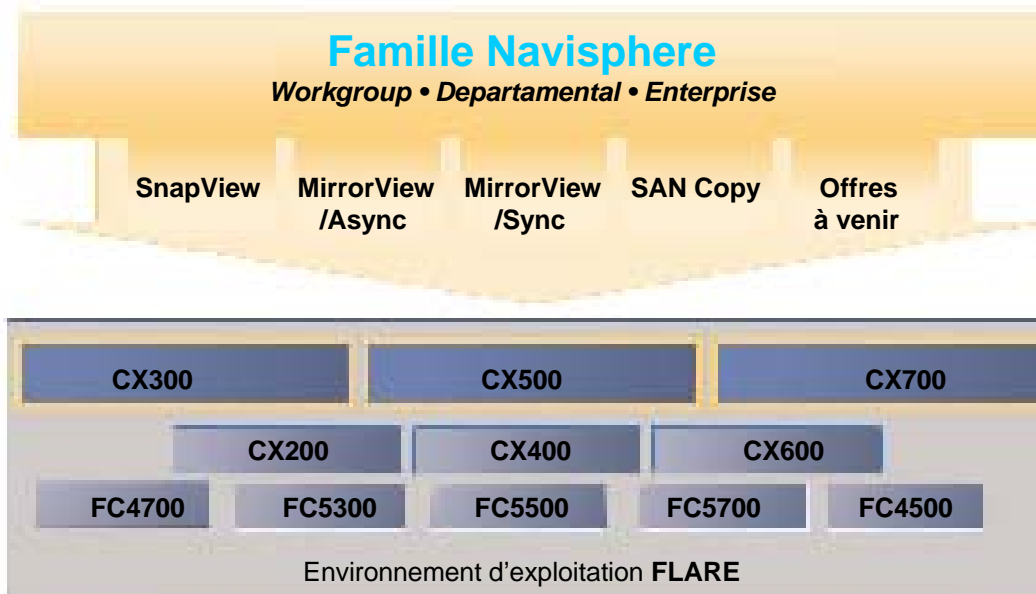
## ■ Le vault

- Espace réservé sur les 5 premiers disques
- Configuré en RAID3
- Hébergement du Système d'exploitation (Flare Code)
- Espace persistant pour les données des caches

# Baie disque : CX300

- Redondance totale :
  - Alimentations
  - Ventilateurs
  - Storage processeurs
  - Batterie de secours (SPS)
  - Disques double accès (SPA et SPB)
  - Caches (via les miroirs)
  - Liens fibres vers les serveurs
- Disque Hot Spare
- Garantie 3 ans 24h24 7j7 sous 4H

# Baie disque : Administration



- Suite de management
- Basée sur du web (java)
- Echanges sécurisés SSL
- Configuration de la baie
- Gestion de la baie
- Suivi des incidents
- Gestion applications optionnelles

# Baie disque : Administration

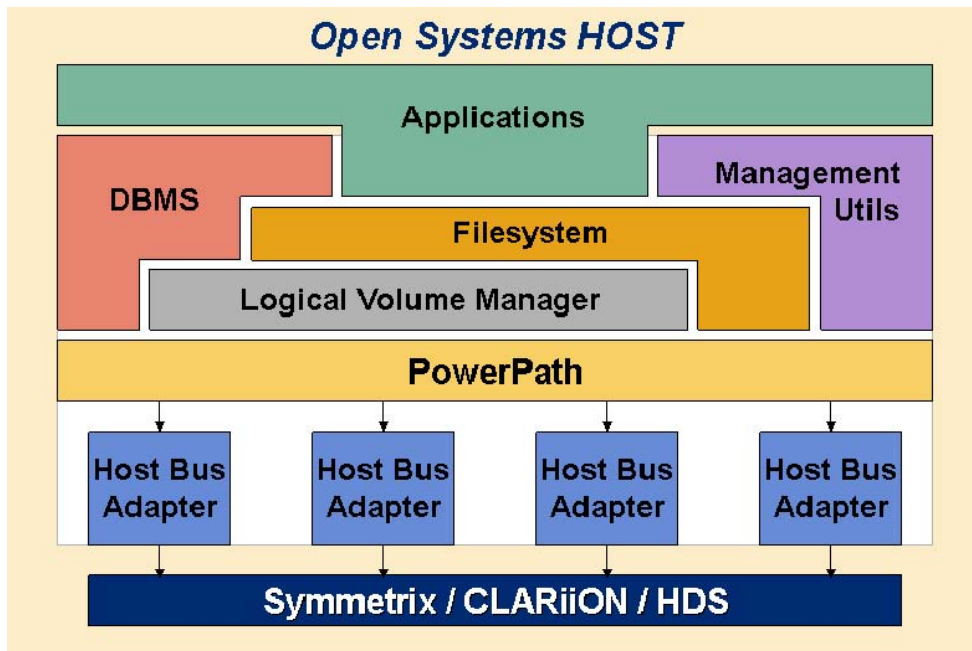
- Fonctionnalités de base :
  - Configuration des LUN et des groupes RAID
  - Masquage de LUN, contrôle accès SAN distribué
  - Modification dynamique de la configuration
  - MetaLUN, VirtualLUN
  - Navisphere CLI (interface ligne de commande)
- Applications optionnelles
  - MirrorView (mise en miroir entre plusieurs baies)
  - Snapview (capture d'une image instantanée d'un LUN)
  - SanCopy (copie des données entre baies)
  - Analyser (outil d'analyse de performances)



# Serveurs : PowerEdge 2850

- Caractéristiques :
  - Bi-processeurs Xéon 3.2 GHZ, 1 Mo de cache
  - 2 Go de mémoire DDR2 à 400 Mhz
  - Bus système à 800 MHZ
  - Contrôleur RAID, SCSI ultra 320, 6 disques
  - 2 interfaces Gigabit Ethernet
  - 2 cartes Qlogic 2340
  - Linux Red Hat Advanced Server 4
- Redondance :
  - Alimentations électriques
  - Ventilateurs
  - Barrette de mémoire spare (Memory Spare Bank)
  - Cartes Qlogic
  - Disques (Raid 1)
- Garantie 3 ans 24h24 7j7 sous 4H

# Serveurs : POWERPATH

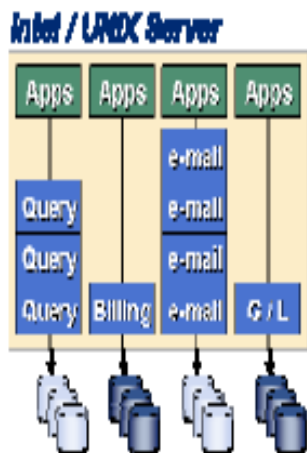


- Logiciel résidant sur les serveurs
- Amélioration des performances
- Amélioration de la disponibilité
- Basculement des trajets
- Equilibrage de la charge IO
- De 2 à 32 canaux

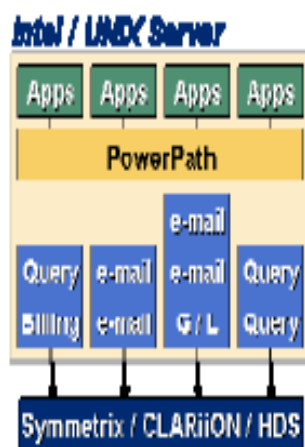
# Serveurs : POWERPATH

## Fonctionnalité d'équilibrage de charge

Sans PowerPath

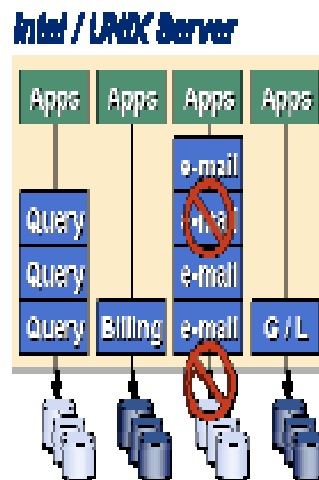


Avec PowerPath

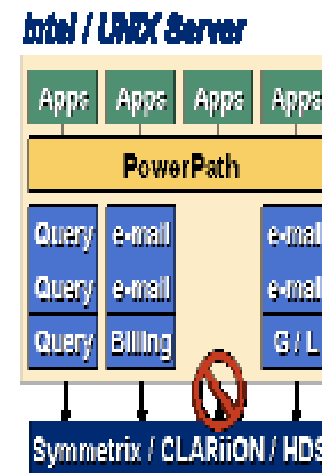


## Fonctionnalité de failover automatique

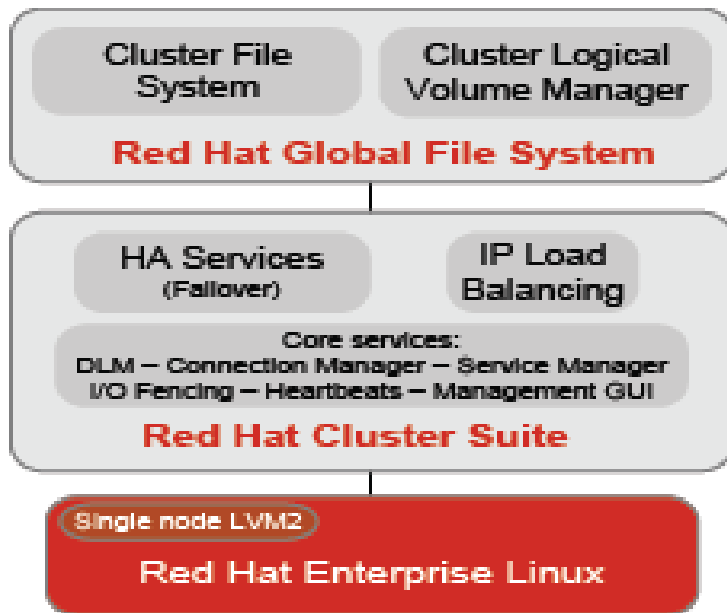
Sans PowerPath



Avec PowerPath



# Serveurs : Cluster Suite

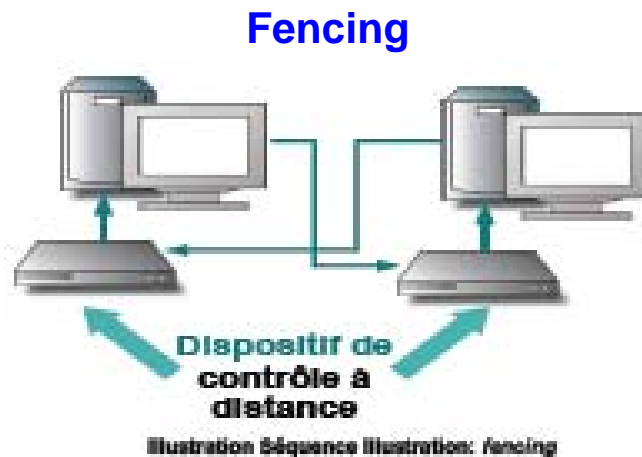


- Logiciel de clustering
- Installé au-dessus de Linux Red hat
- Sous-ensemble de GFS (Global File System)
- Cluster Manager (haute disponibilité)
- Equilibrage de charge IP

# Serveurs : Cluster Suite

## ■ Cluster Manager :

- Gestionnaire de grappe fonctionnant en mode actif/passif ou actif/actif
- Jusqu'à 16 nœuds
- Garantie complète de l'intégrité des données (fencing)
- Outil de gestion graphique
- Concept de service identifié par un nom et une adresse IP




- Garantir l'intégrité des données
- Sortir Physiquement un nœud du cluster
- Dispositif de contrôle à distance
- Extinction du serveur via le réseau

# Sauvegarde : Powervault 123T

- Caractéristiques :
  - Bibliothèque LTO-3
  - 24 cartouches, 1 ou 2 lecteurs
  - SCSI ultra 320 ou interface Fiber Channel
  - 9,6 To/19,2 To, 400Go/800Go
  - 576 Go/heure/1,15 To/heure
  - Lecteur de code barres
- Garantie 3 ans 24h24 7j7 sous 4H
- Logiciel :
  - Time Navigator

# Configuration mise en place

- CX300
  - 9 disques de 300 Go pour les données
  - 1 disque Hot Spare
  - 2 Raid Group (5 et 4 disques) en Raid 5 (environ 2 To utile)
  - 6 LUN (3 par Raid Group et par SP)
- Serveurs
  - 6 volumes logiques, 3 par serveur
  - Accéder via les 4 liens Fiber Channel
- Cluster
  - Actif/Actif
  - Samba, NFS (3 volumes par serveur), CUPS



# Plan de l'exposé

- Contexte
- Approches
- Technologie SAN
- Technologie de clustering
- Solution mise en œuvre
- Bilan



# Bilan

- Solution en place depuis 3 mois
- Très performante et très sécurisée
- Assistance très efficace (24x7 illimitée)
- Complexe à mettre en œuvre
- Investissement important pour maîtriser la technologie
- Package PowerPath pas inclus dans RHAS
- Peu de messages générés par Cluster Suite