

Stockage Réseau

Le stockage s'échappe du système
pour devenir une fonction réseau

Philippe Latu / Université Toulouse III - Paul Sabatier / www.inetdoc.net

[Philippe.latu\(at\)inetdoc.net](mailto:Philippe.latu(at)inetdoc.net)

■ Les enjeux

- Besoins en progression constante
- Migration DAS vers (SAN | NAS)
- Hétérogénéité et Interopérabilité
- Continuité de service

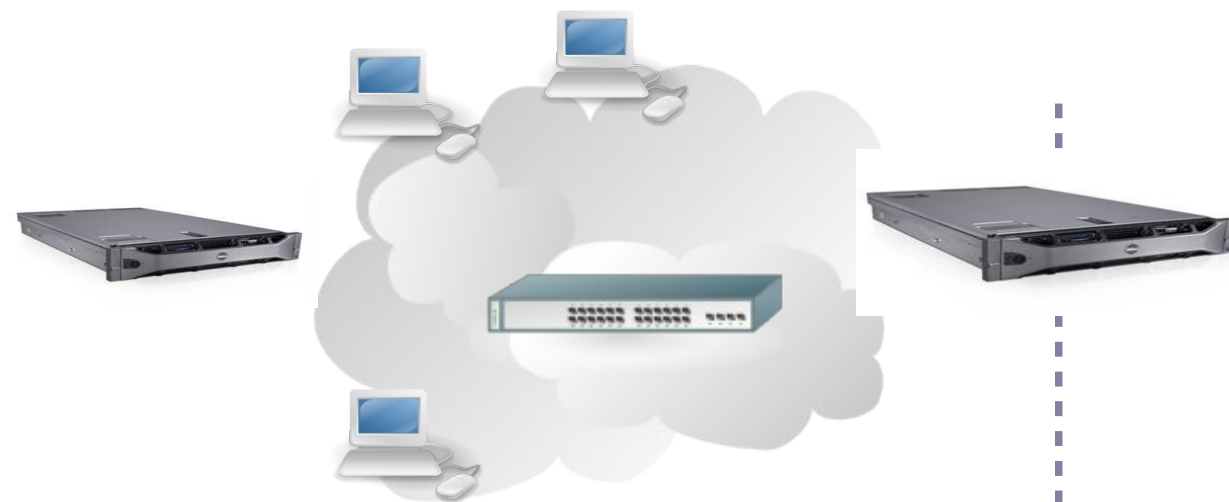
■ Les termes

- Manageability
- Availability
- Scalability

■ Les acronymes

- DAS : Direct Attached Storage
- SAN : Storage Area Network
- NAS : Network Attached Storage

■ Différences entre modes d'accès : fichier ou bloc



Réseau IP «frontal»

- Hôte ↔ hôte
- Application ↔ système de fichiers
- Client ↔ Serveur
- NFS / CIFS
- NAS



Réseau de stockage «dorsal»

- Hôte ↔ stockage
- Système de fichiers ↔ périphérique
- Application ↔ périphérique
- Virtual FS / Ext4 / NTFS
- SAS / SATA / SCSI / PATA
- SAN

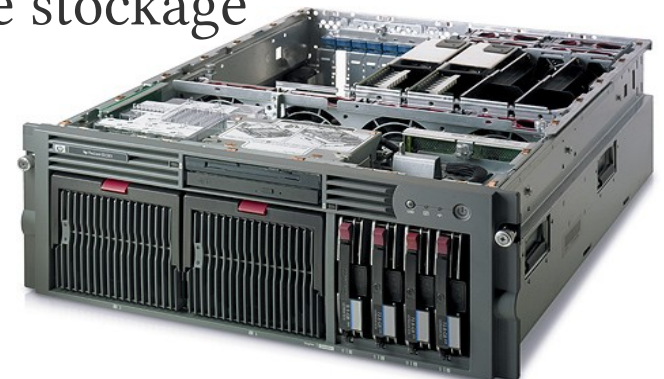
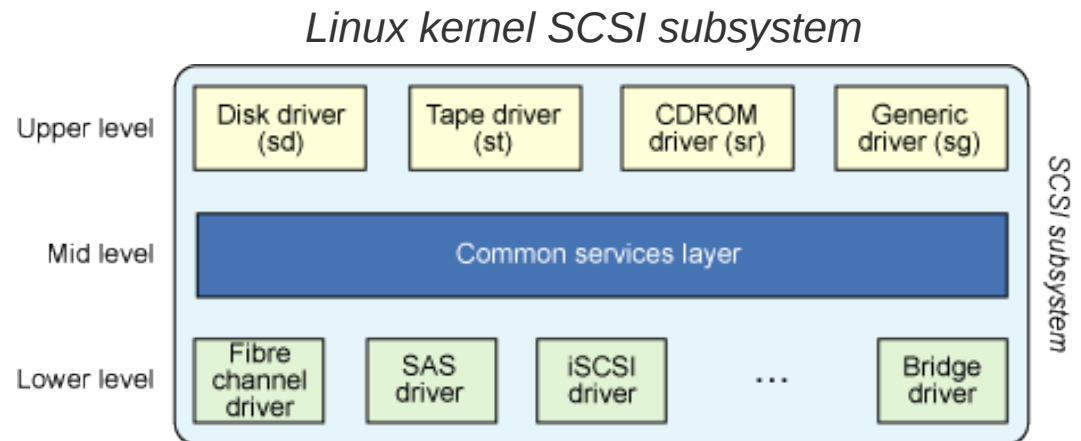
■ Caractéristiques

■ Évolutions

- SCSI → SAS
- PATA → SATA
- FC (ANSI) → FCOE

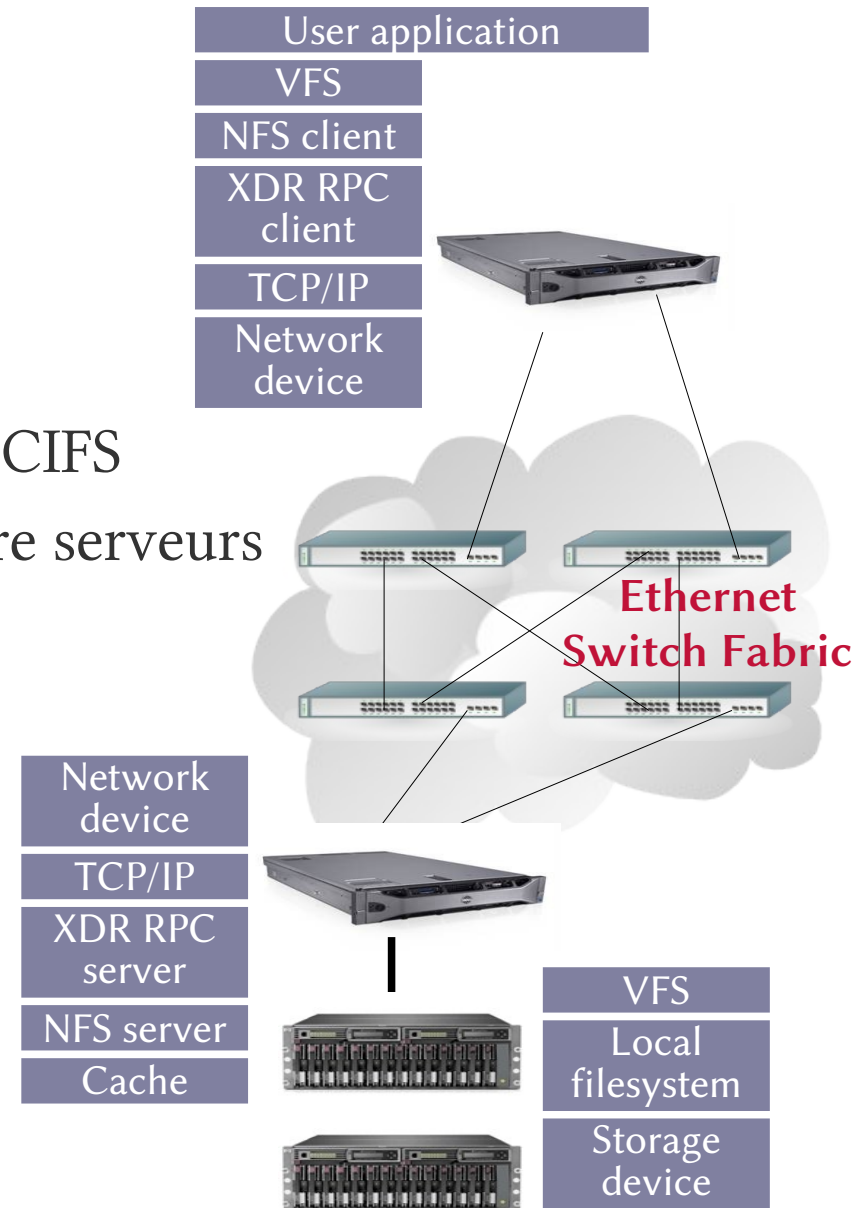
■ Limitations

- Distance entre système et périphériques
- Nombre de disques par châssis
- Partage de périphériques entre systèmes
- Dimensionnement serveur vs. Capacité de stockage
- Retour sur investissement (ROI)



■ Caractéristiques

- Hôte connecté au réseau IP
- Accès en «mode fichier»
- Évolutions
 - Systèmes de fichiers réseau NFS ou CIFS
 - Système de fichiers partageable entre serveurs
 - Liens redondants possibles
- Limitations en performances
 - Réseau
 - Système de fichiers
 - Système d'exploitation



■ Caractéristiques

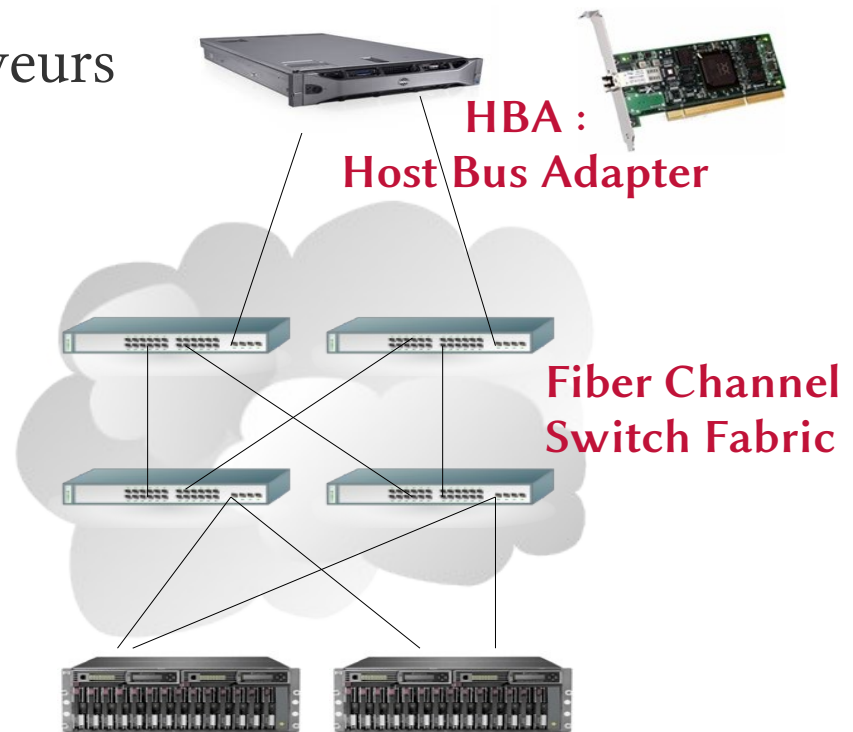
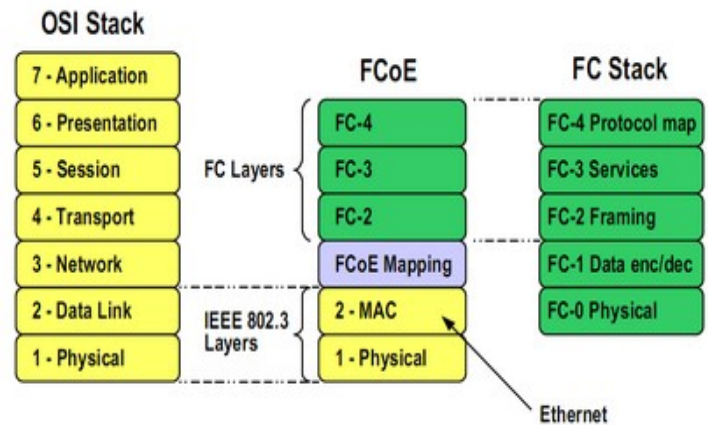
- Hôte connecté à un commutateur
- Accès en «mode bloc»

■ Évolutions

- Fiber Channel over Ethernet (FCoE)
- Sous-système partageable entre serveurs
- Liens redondants possibles

■ Limitations en performances

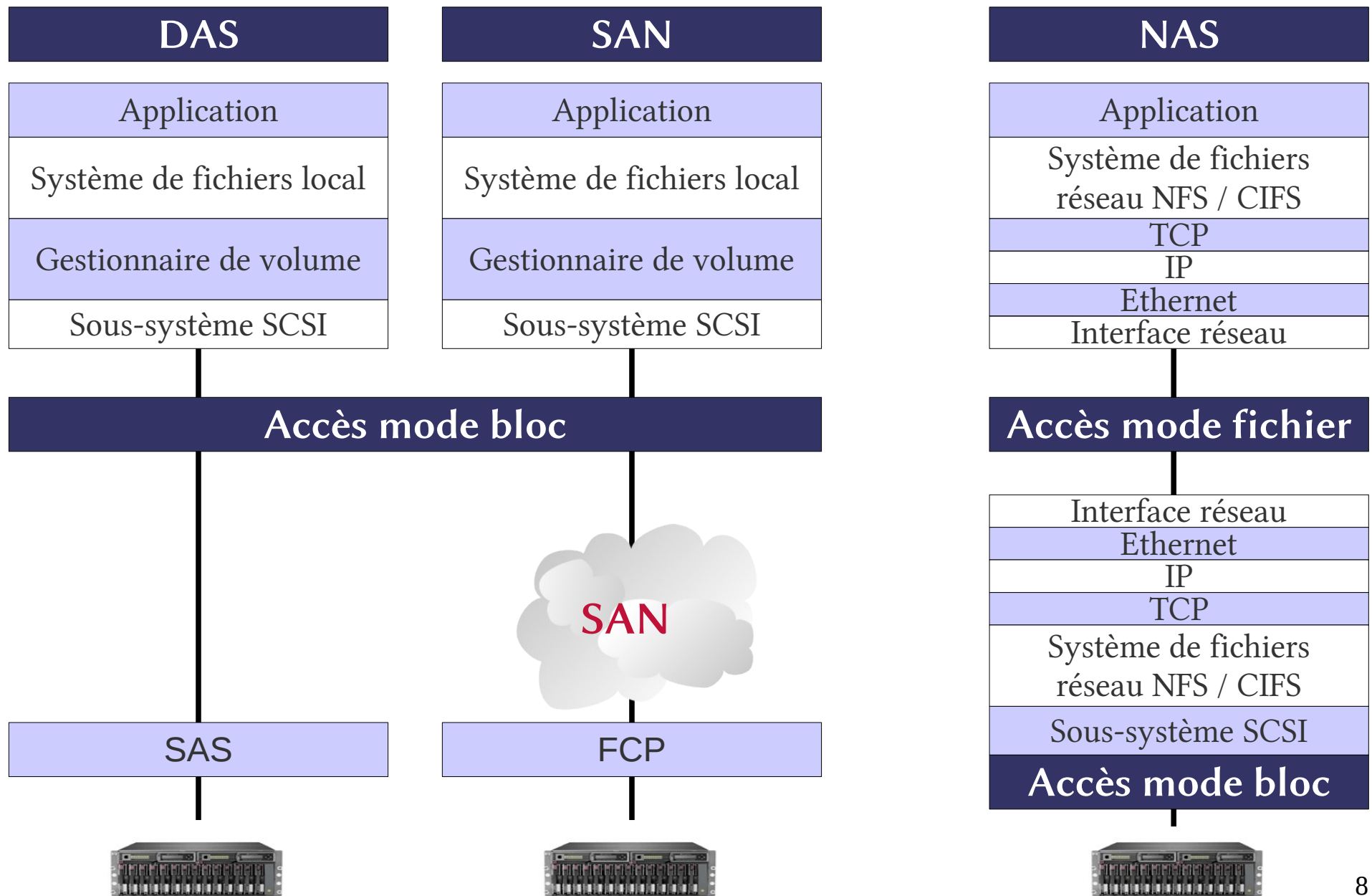
- Coût du port de commutateur
- Choix d'interfaces (HBA/FC) limité



■ Caractéristiques système

	DAS	NAS	SAN
Accès	Mode bloc	Mode fichier	Mode bloc
Connexion	Série – SAS Parallèle - SCSI	Ethernet	Fiber Channel
Performances d'accès	Très bonnes	Moins bonnes	Très bonnes
Limite des performances	Sous-système SCSI du noyau	Système de fichiers NFS / CIFS	Commutation Fiber Channel
Augmentation de capacité	Arrêt du système obligatoire	Très facile	Complexe suivant l'architecture
Évolutivité et Continuité d'exploitation	Faible	Moyenne	Élevée

■ Caractéristiques réseau



■ Caractéristiques

■ Accès en «mode bloc» sur lien Ethernet

■ Évolutions

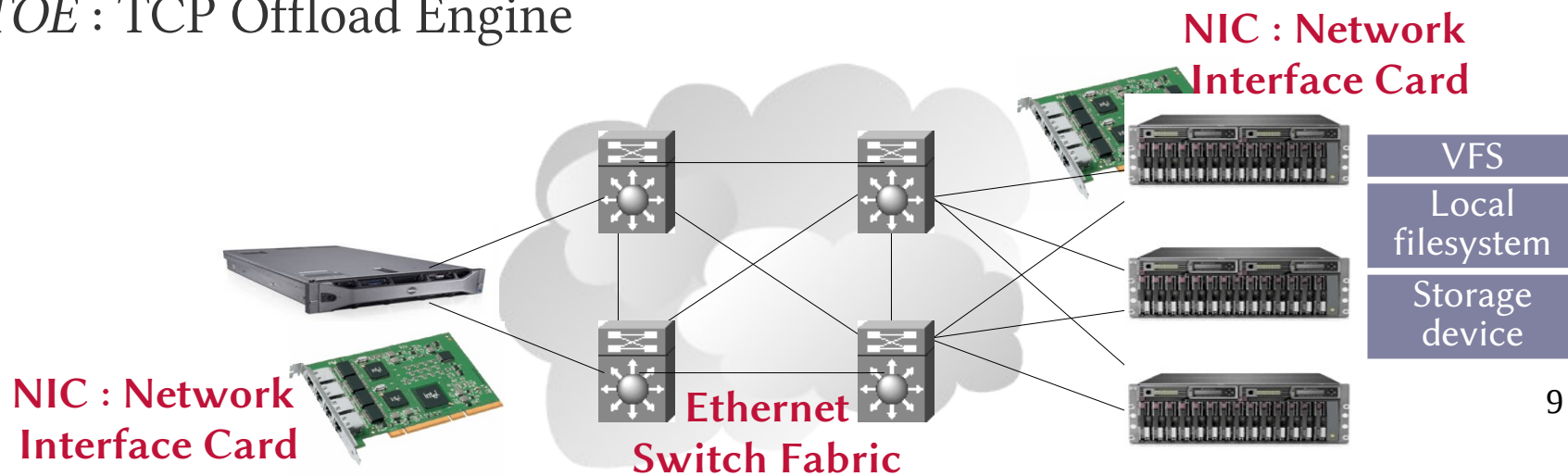
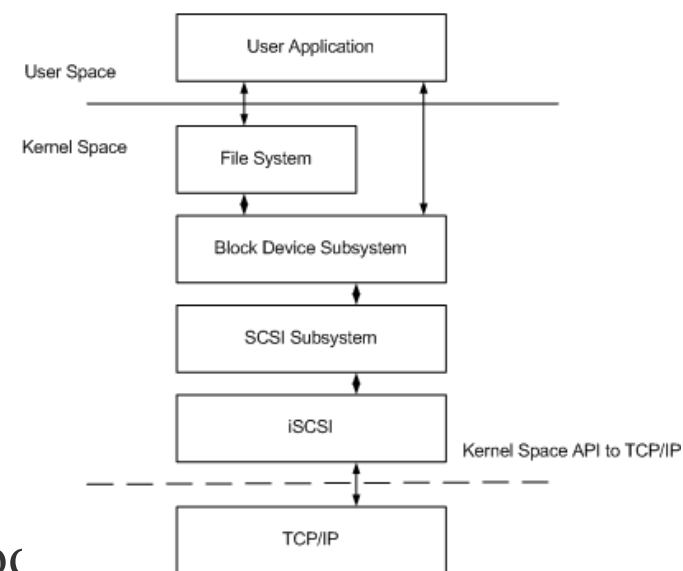
- Algorithmes TCP adaptés (HTCP)
- Balance de charge multi-liens (LACP)

■ Limitations

- Performances réseau
- Conflits entre fonctions TCP accès «mode bloc»

■ Termes

- *Initiators* : HBAs ou NICs côté serveurs (maîtres)
- *Targets / LUNs* : HBAs ou NICs côté stockage (esclaves)
- *TOE* : TCP Offload Engine



■ Caractéristiques

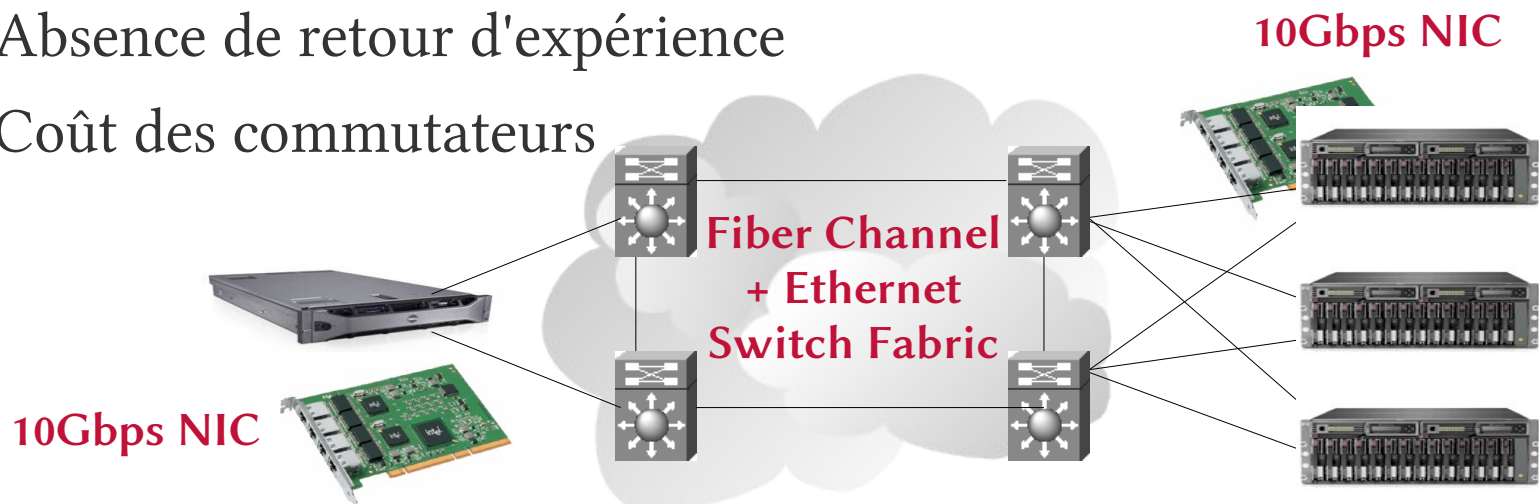
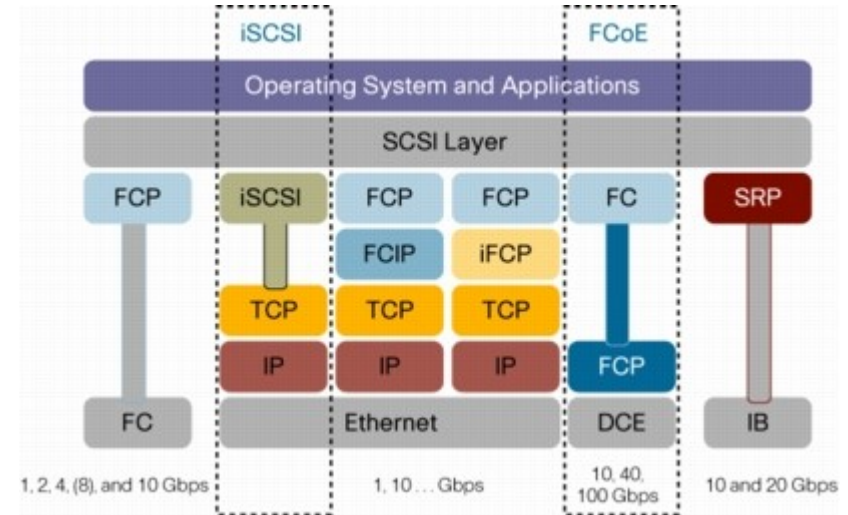
- Accès en «mode bloc» sur réseau IP

■ Évolutions

- Commutation unifiée
 - LAN + SAN
- Interfaces 10Gbps unifiées
 - réseau + stockage

■ Limitations

- Spécifications trop récentes
- Absence de retour d'expérience
- Coût des commutateurs



■ Contraintes

■ Reprise de service en cas de panne

- Catastrophes naturelles
- Erreurs humaines

■ Sauvegarde

- Opérations de maintenance
- Pannes et défauts matériels

■ Accès multi-liens

- Redondance
- Balance de charge
- Qualité de service (QoS)

■ Réplication

- Disponibilité
- Sauvegarde

■ *Redundant Array of Independent Disks (RAID)*

■ Deux types d'implémentation

- Logicielle – sous-système «*device manager*» du noyau Linux
- Matérielle – carte contrôleur avec un système propre «*firmware*»

Niveau RAID	Description	Nombre minimum de disques	Capacité utile (nombre de disques)
0	<i>Striping</i> / Concaténation	2	N
1	Miroir	2	N/2
1 + 0	Miroir puis <i>Striping</i> / Concaténation	4	N/2
5	<i>Stripes</i> avec parité distribuée et E/S aléatoires	3	N - 1
6	<i>Stripes</i> avec deux calculs de parité différents distribués et E/S aléatoires	4	N - 2

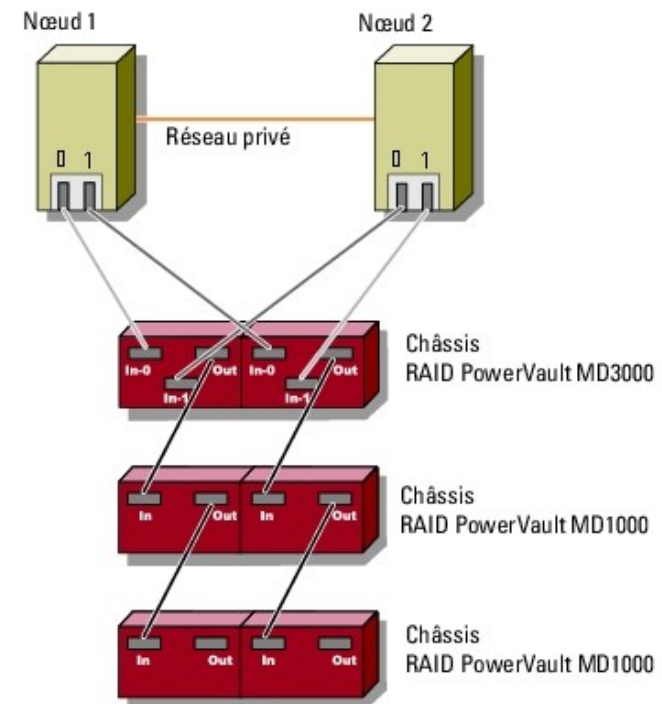
■ Entrées/Sorties redondantes - *Multipath I/O*

■ Tolérance aux pannes dynamique

- *Failover / Recovery*
- Optimisation du coût d'administration

■ Choix entre deux modes

- Actif / Passif
 - Tolérance aux pannes
 - Détection d'erreur automatique
- Actif / Actif
 - Augmentation de performances
 - Augmentation des débits



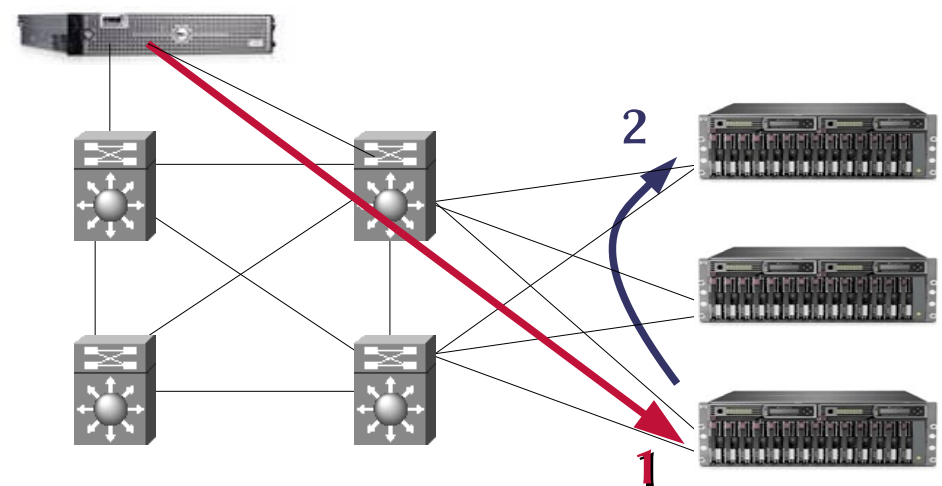
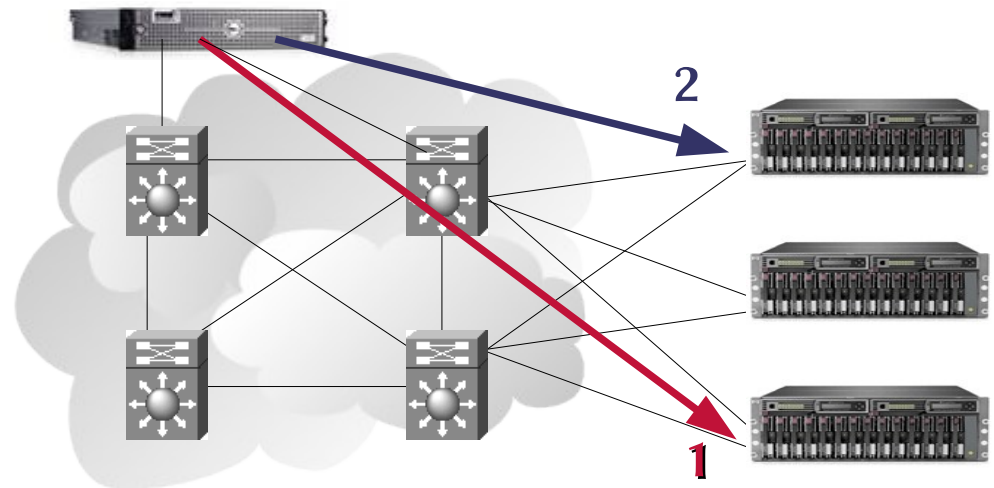
■ Deux modèles distincts

■ Niveau système ou noyau

- Synchrone ou asynchrone
- Pilotage au niveau serveur
 - LVM | cron | rsync

■ Niveau sous-système

- Synchrone ou asynchrone
- Pilotage au niveau contrôleur
 - *Mirroring*
 - Performances réseau imposées
 - Distances limitées



■ Types de sauvegarde

■ Complète

- Volume ou système de fichiers complet
- Temps et consommation de bande passante très importants

■ Incrémentale

- Différence depuis la dernière sauvegarde
- Temps et consommation de bande passante peu importants

■ Différentielle

- Différence depuis la dernière sauvegarde complète
- Temps et consommation de bande passante moins importants

■ Hors ligne

- Fenêtre de blocage des écritures imposée
- Impact important sur l'architecture des services

■ En ligne

- Contrôle d'intégrité difficile
- Conflits potentiels entre services et sauvegarde

- Types d'opérations
 - Ajout/Retrait d'unités de disques
 - Maintenance avec utilisation temporaire de disques
 - Augmentation/Diminution de la capacité de stockage
 - Transferts entre volumes logiques sur un même système
 - Redimensionnement dynamique
 - Extension d'un système de fichiers en ligne
 - Déplacements de données entre unités de disque
 - Préparation à l'extraction d'unités de disque
 - *Snapshots*
 - Copie instantanée de l'état d'un volume logique
 - Réplication
 - Copie entre volumes logiques

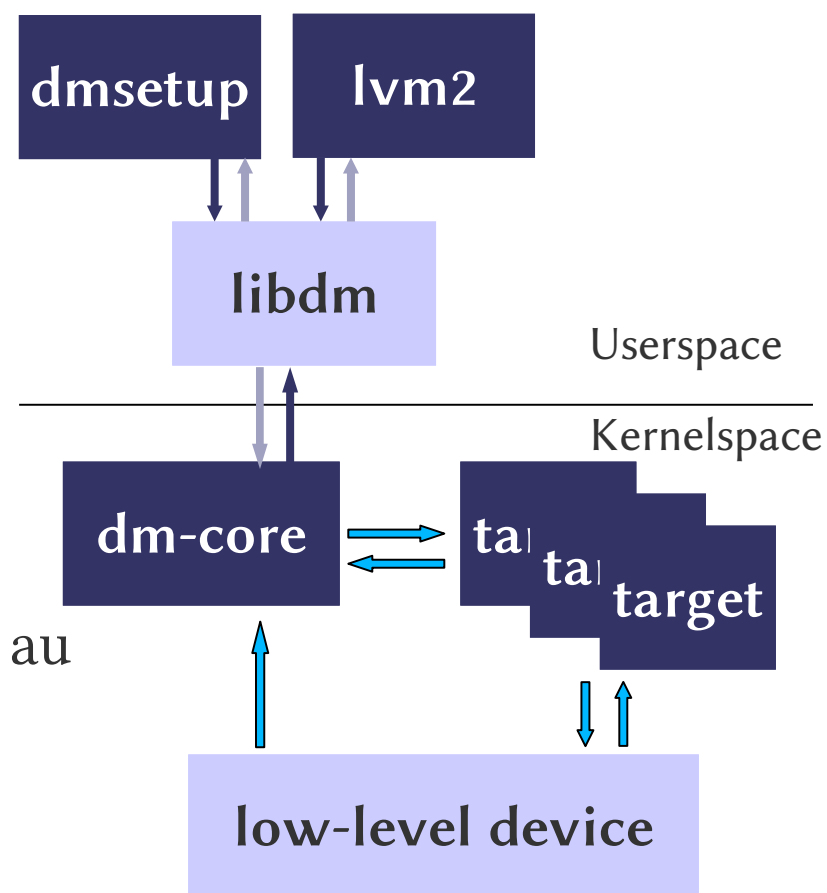
■ *Device mapper*

■ Cartographie des requêtes sur les unités de disque

- Redirection
- Mise en attente
- Chiffrement
- Gestion de lien

■ Gestionnaire de périphériques en mode bloc

- Ajout/Retrait d'unités en mode bloc au dessus de périphériques de stockage physiques



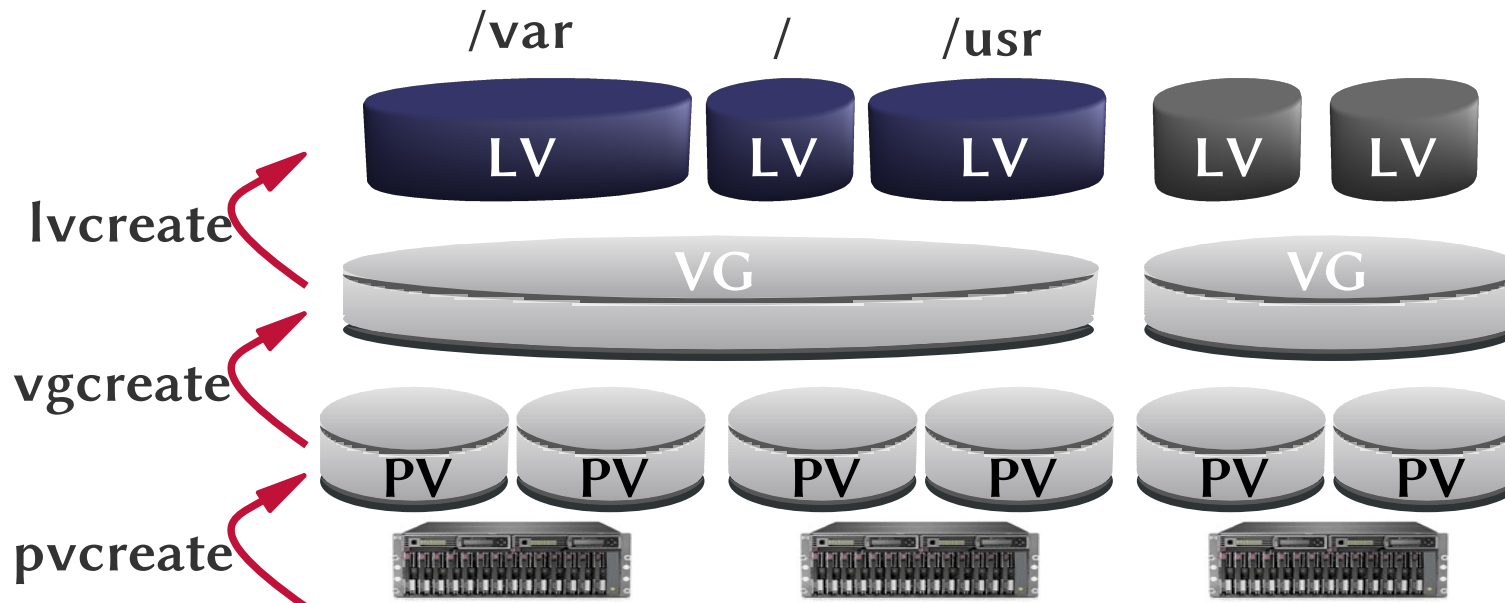
■ *Logical Volume Manager*

- Gestionnaire de périphérique mode bloc au niveau système
 - Partitions d'unités de disque
 - Unités SAS/SATA/PATA
 - LUNs iSCSI
 - Réseau FCoE
- **Vue système homogène**
 - N Périphériques physiques vus comme un périphérique logique
- **Analogie entre volume et partition**
 - Formatage et création d'un système de fichiers
 - Partition d'échange (*swap*)
 - Accès directs depuis un gestionnaire de bases de données
- **Changements dynamiques de configuration**

■ Linux LVM2

- Espace noyau
 - *Device mapper modules*
- Espace utilisateur
 - Paquet lvm2
- *Striping en option*

```
-[-] Multiple devices driver support (RAID and LVM)
< > RAID support
<M> Device mapper support
[ ] Device mapper debugging support
<M> Crypt target support
<M> Snapshot target
<M> Mirror target
<M> Zero target
<M> Multipath target
<M> I/O delaying target (EXPERIMENTAL)
[*] DM uevents (EXPERIMENTAL)
```



■ Réplication synchrone

■ RAID1 logiciel entre DAS et SAN

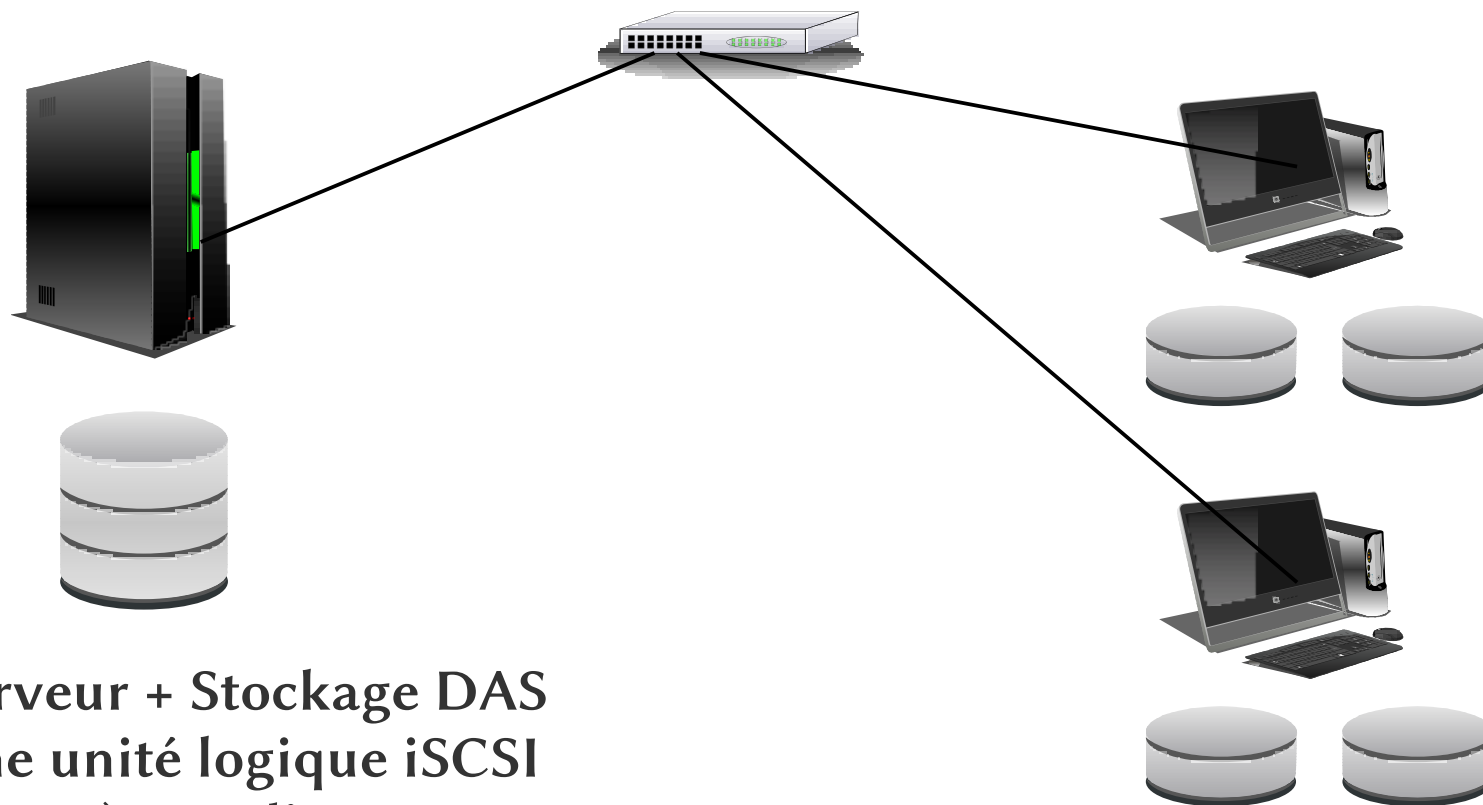
- Une unité de disque locale
- Une unité de disque iSCSI
- RAID1 entre les deux unités

■ Réplication asynchrone

■ *Snapshot* LVM entre DAS et SAN

- Une unité de disque locale
- Une unité de disque iSCSI
- *Snapshots* périodiques entre les deux unités

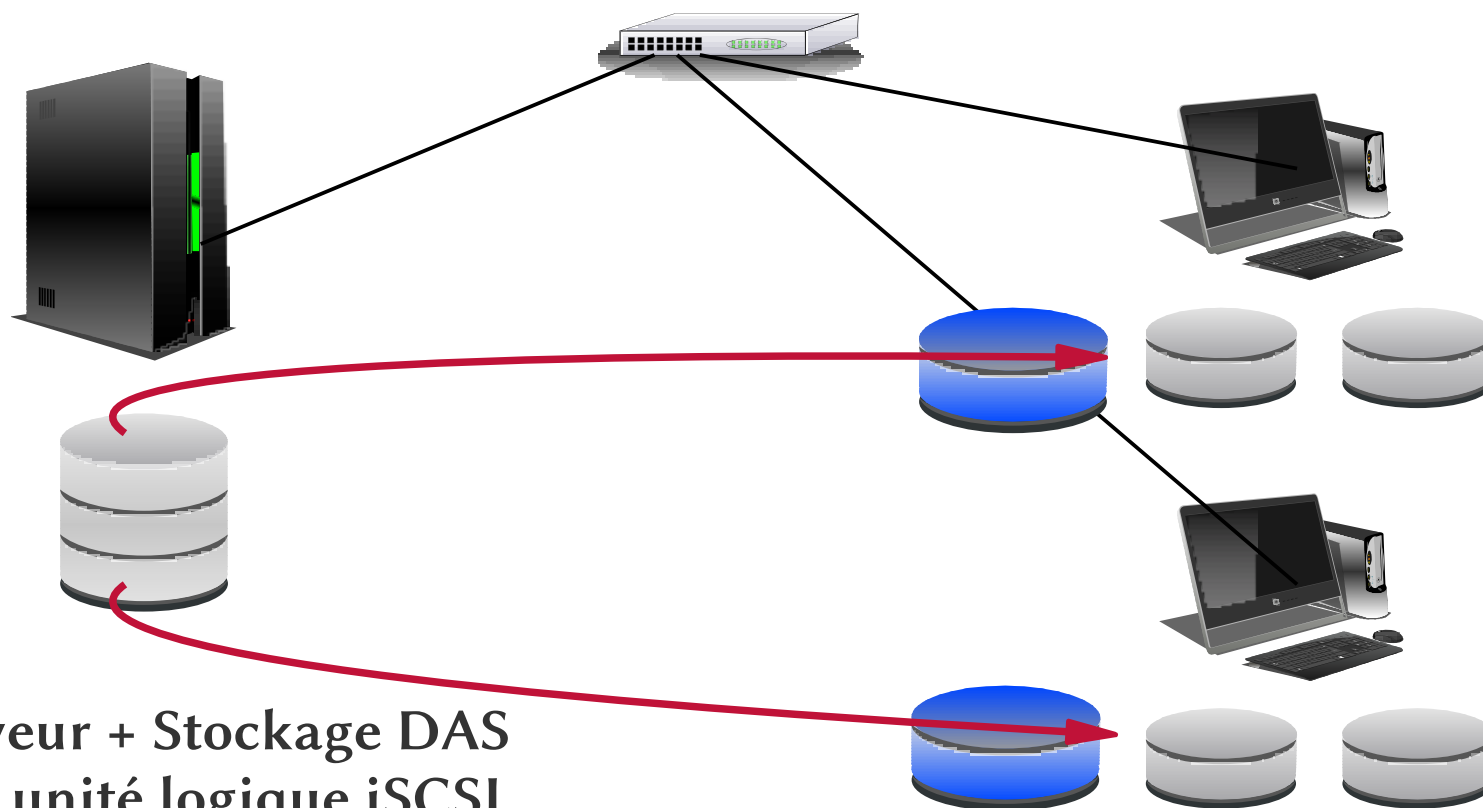
■ Réplication synchrone – phase 0



- Serveur + Stockage DAS
- Une unité logique iSCSI (target) par client

- Client + Stockage DAS
- deux unités physiques par poste

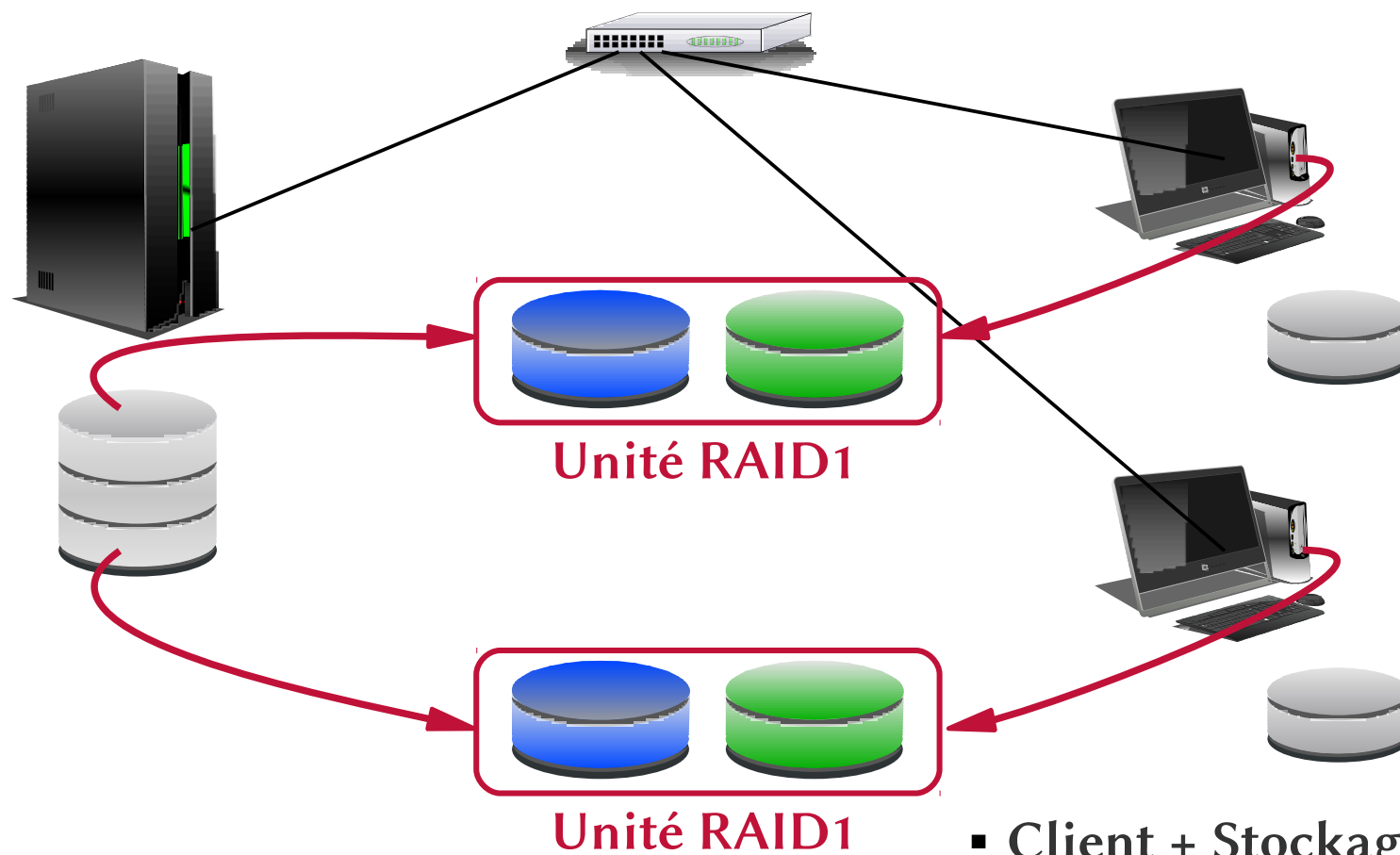
■ Réplication synchrone – phase 1



- Serveur + Stockage DAS
- Une unité logique iSCSI (target) par client

- Client + Stockage DAS
- deux unités physiques par poste
- une unité logique iSCSI (initiator)

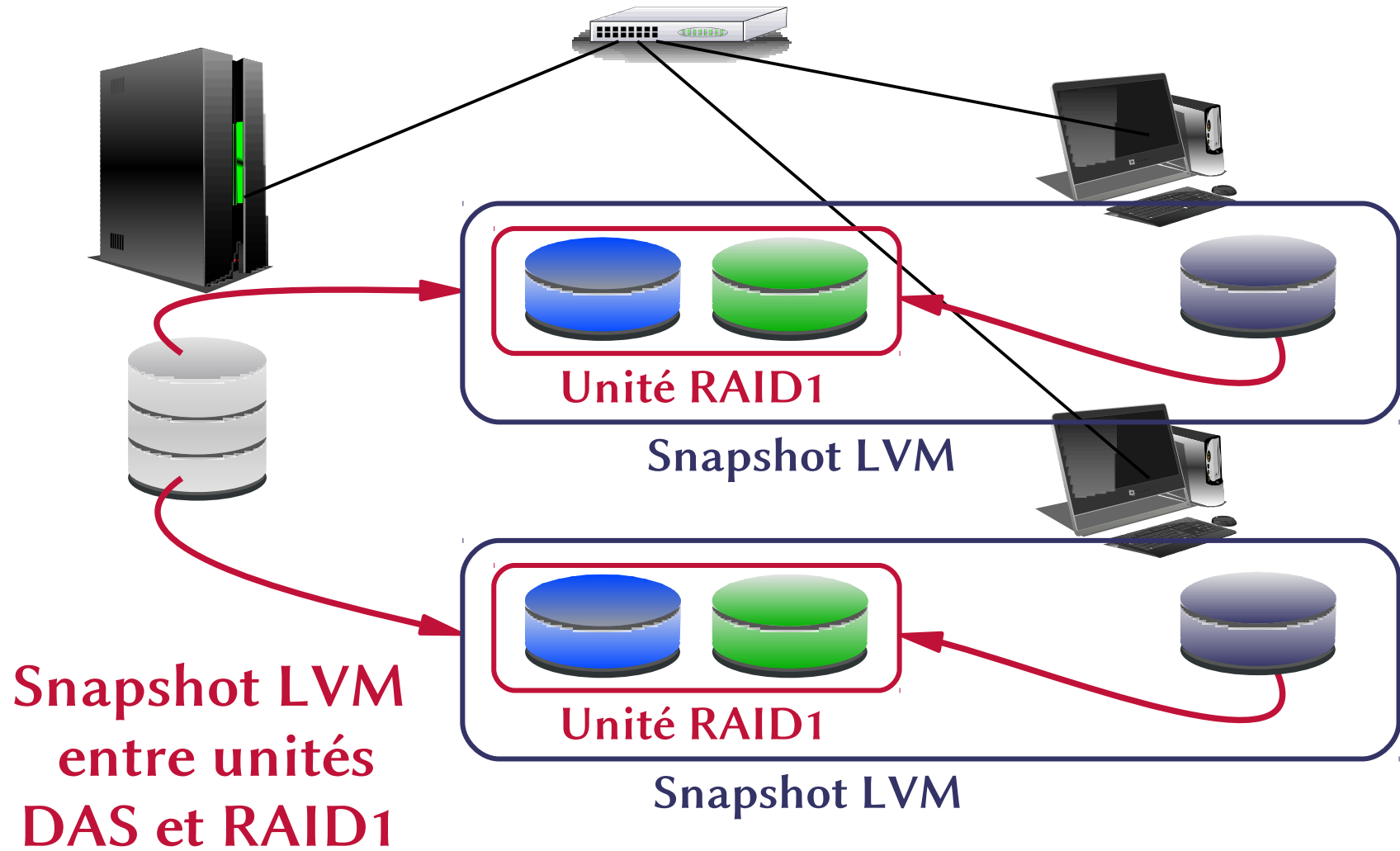
■ Réplication synchrone – phase 2



RAID1 avec DAS + SAN
=
Réplication synchrone

- Client + Stockage DAS
- une unité physique par poste
- une unité logique RAID1

■ Réplication asynchrone – phase 3



=

Réplication asynchrone

■ Travaux pratiques

- Introduction au réseau de stockage iSCSI
- <http://www.linux-france.org/prj/inetdoc/cours/admin.reseau.iscsi/>

■ Technologies

- NAS : http://fr.wikipedia.org/wiki/Stockage_en_r%C3%A9seau_NAS
- SAN : http://fr.wikipedia.org/wiki/Storage_Area_Network
- FcoE : <http://fr.wikipedia.org/wiki/FCoE>
- iSCSI : <http://fr.wikipedia.org/wiki/ISCSI>
- LVM : http://fr.wikipedia.org/wiki/Gestion_par_volumes_logiques

■ Documentation

- LVM : <http://tldp.org/HOWTO/LVM-HOWTO/>
- Wiki : <http://sources.redhat.com/lvm2/wiki/>