

Stockage des données

Octobre 2002

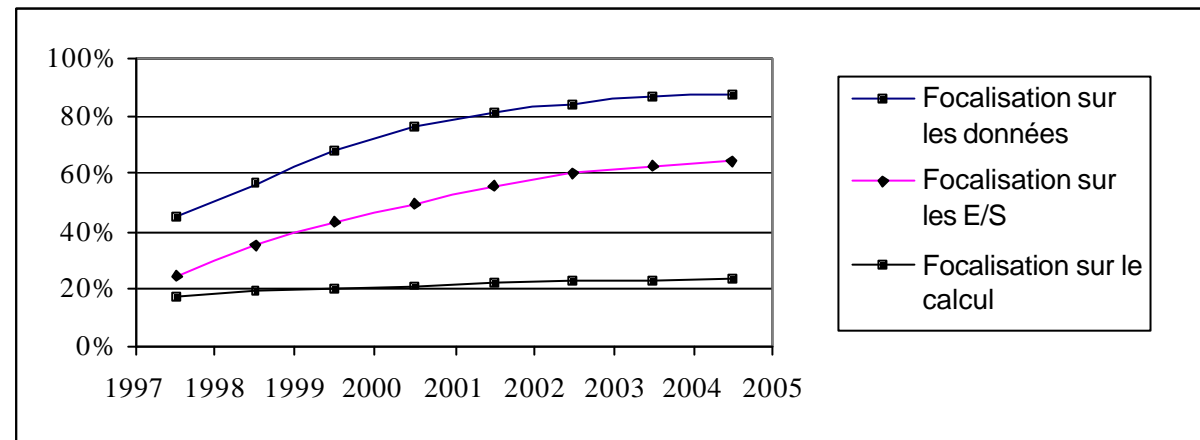
René J. Chevance

- Quelques données utiles
- Évolution des disques magnétiques
- Problématique du stockage
- Niveaux fonctionnels - Systèmes de fichiers
- Accès aux fichiers distants - NFS, CIFS et DAFS
- Organisation des disques - JBOD, SBOD et RAID
- Virtualisation du stockage
- Architectures de stockage
DAS, SAN; NAS et iSCSI
- Modèle d'architecture SNIA
- Management du stockage
- Exemples de solutions de stockage

Version provisoire

Les différents schémas sont tirés d'un document destiné à la publication dans un ouvrage anglais. Les figures comportent donc des termes anglais.

■ Évolution de la part du stockage dans les dépenses en matière de serveurs (Source GartnerGroup)



- **Focalisation sur les données : conservation de la quasi-totalité des données «en ligne»**
- **Focalisation sur les entrées-sorties : le temps d'accès aux données est une dimension critique (OLTP)**
- **Focalisation sur le calcul : les caractéristiques du stockage des données ne sont pas critiques**

■ Définition de quelques grandeurs en matière de stockage

Nom	Abréviation	Valeur
Gigaoctet	Go	10^9 octets
Téraoctet	To	10^{12} octets
Pétaoctet	Po	10^{15} octets
Exaoctet	Eo	10^{18} octets

Quelques données utiles(2)

- Production mondiale de contenus originaux en 1999 en To - Étude « How much information? » de l'Université de Berkeley

Média	Type de contenu	To/an Estimation haute	To/an Estimation basse	Taux de croissance %
Papier	Livre	8	1	2
	Journaux	25	2	-2
	Périodiques	12	1	2
	Documents de bureau	195	19	2
	Sous-total	240000	23	2
Film	Photographies	410000	41000	5
	Cinéma	16	16	3
	Radiographies	17200	17200	2
	Sous-total	427216	58216	4
Optique	CD musique	58	6	3
	CD données	3	3	2
	DVD	22	22	100
	Sous-total	83	31	70
Magnétique	Bande vidéo	300000	300000	5
	Disques PC	766000	7660	100
	Serveurs département.	460000	161000	100
	Serveurs d'entreprise	167000	108550	100
	Sous-total	1693000	577210	55
Total		2120539	635480	50

Quelques données utiles(3)

■ Commentaires

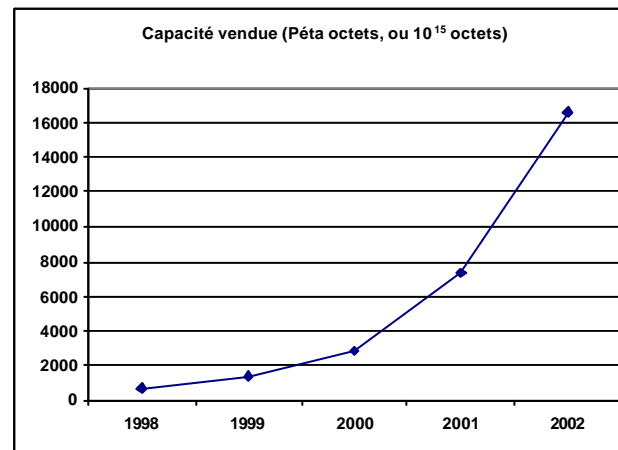
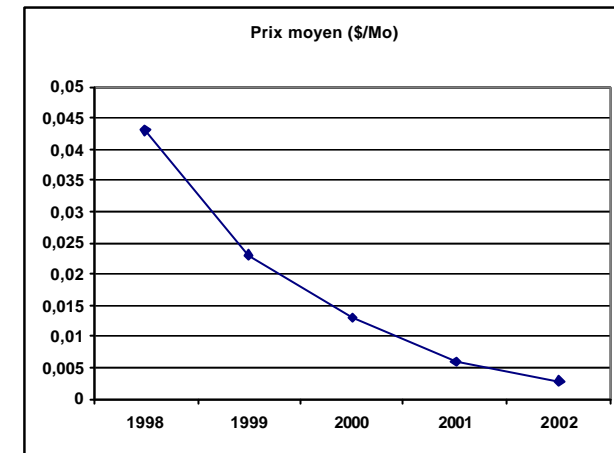
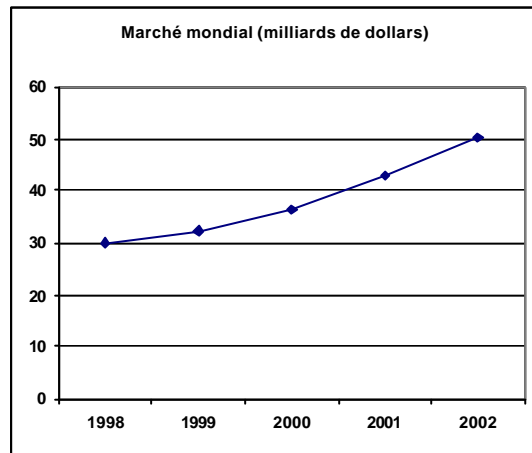
- **La plus grande partie des informations est produite par les individus :**
 - Documents de bureau (80% de l'imprimé)
 - Photographies et radiographies (99%)
 - 55% des disques durs sont installés dans des PCs
 - Web en 2000 = 21 To de pages statiques HTML (2,1 milliards de pages), taux de croissance de 100% par an
- **La quasi totalité de l'information est produite sous forme digitale**

■ Flux de communications aux US en 1999

Média	Teraoctets/an
Radio	788
TV	14150
Téléphone	576000
Poste	150000

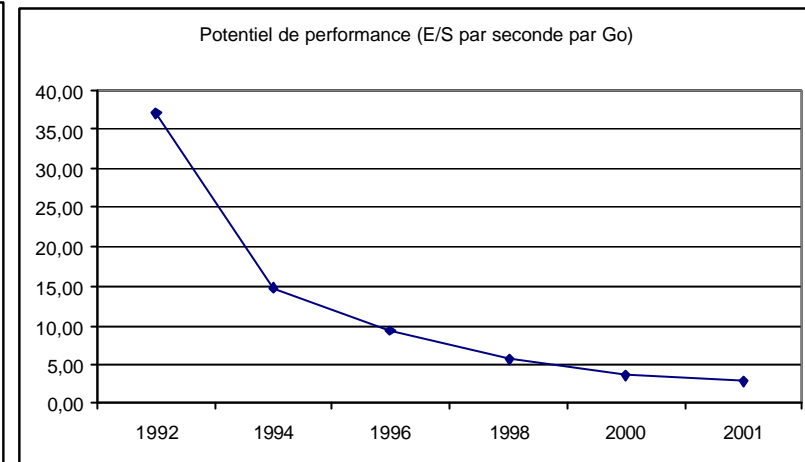
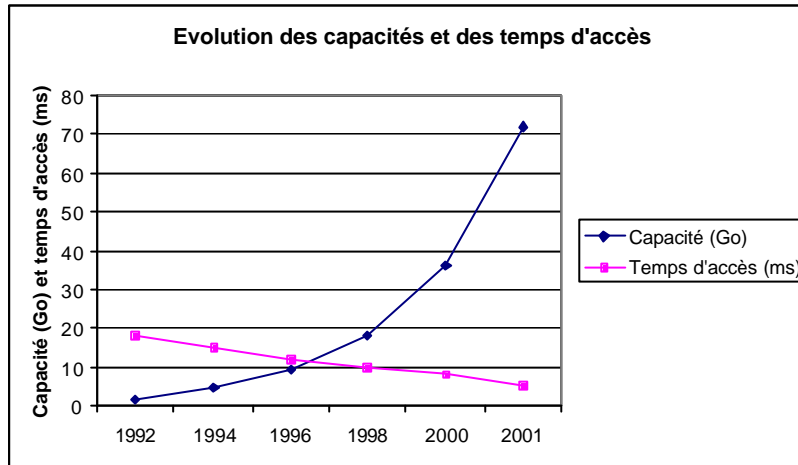
Évolution des disques magnétiques

- **Marché entraîné par le PC**
- **Forte progression des capacités**
- **Progression modérée des temps d'accès (délai rotationnel, positionnement de la tête de lecture, temps de transfert)**



Évolution des disques magnétiques(2)

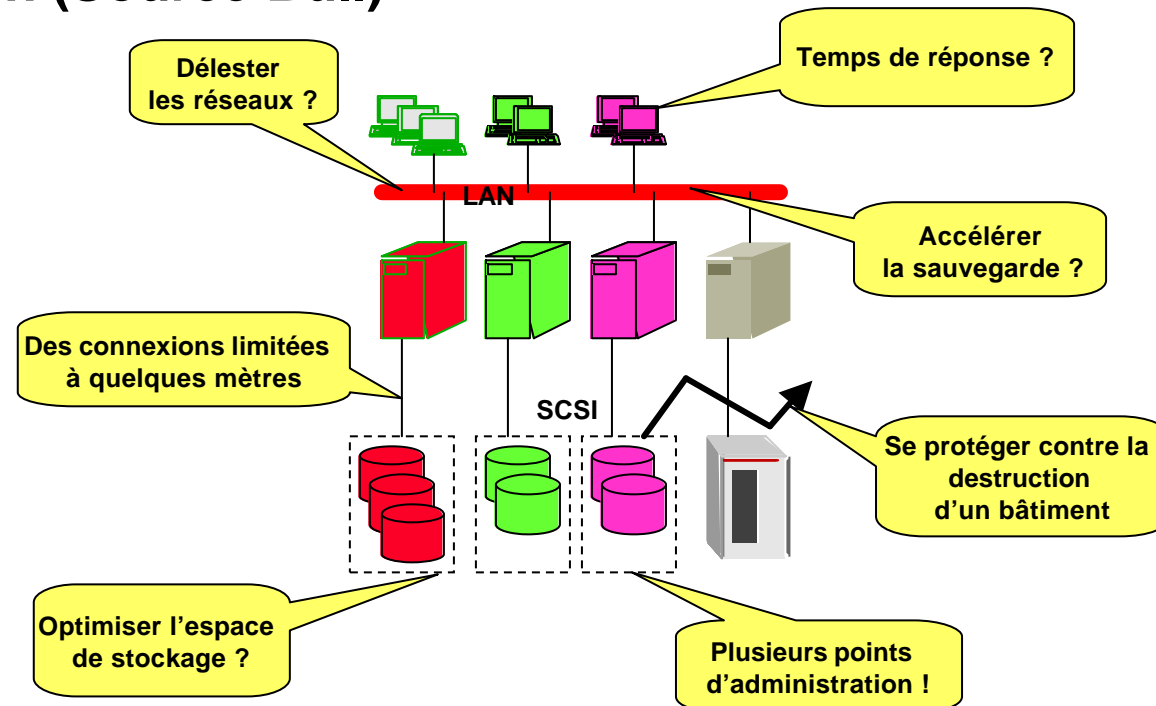
■ Évolution de la capacité et des temps d'accès



- La forte augmentation des capacités et la (relativement) faible progression des temps d'accès entraîne que l'accès aux disques devient, de plus en plus, un facteur critique
- Diminution du nombre d'unités à capacité constante (-> problème de performance) :
 - Répartir les données sur plusieurs disques en parallèle (RAID)
 - Placer les données en mémoire (technique de cache)
 - Écriture en mémoire stable et acquittement rapide (cache sécurisé)
- Vers des disques intelligents? (utilisation des capacités de mémorisation et de traitement contenues dans les unités de disques)

Problématique du stockage

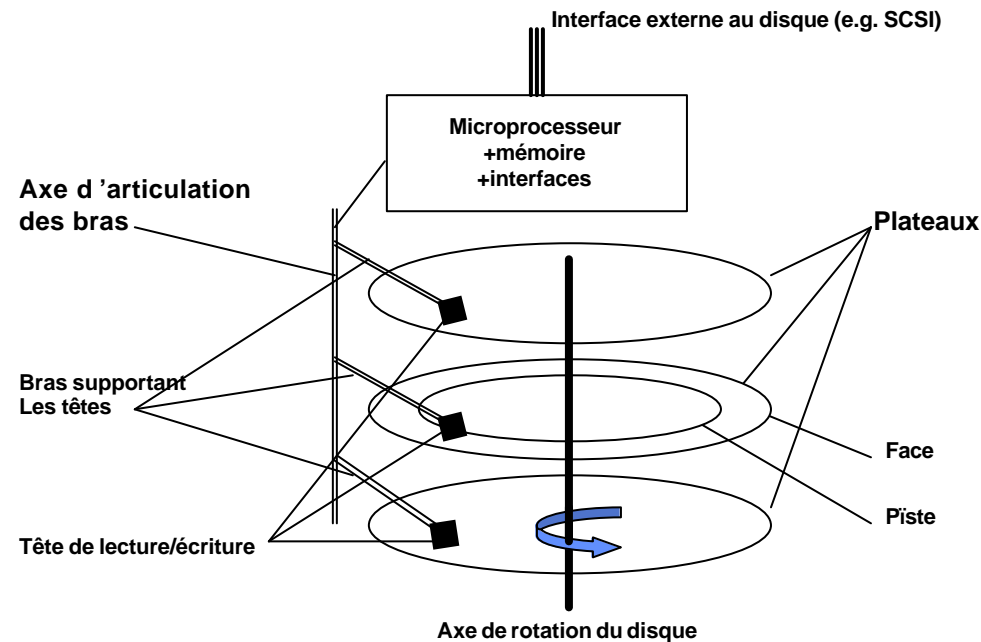
■ Illustration (Source Bull)



- Allègement de la charge sur le réseau local
- Optimisation des temps de réponse
- Diminution/masquage des temps de sauvegarde
- Distance entre serveur(s) et sous-système de stockage
- Protection contre les catastrophes
- Partage et optimisation des ressources de stockage
- Simplification de l'administration

Niveaux fonctionnels Systèmes de fichiers

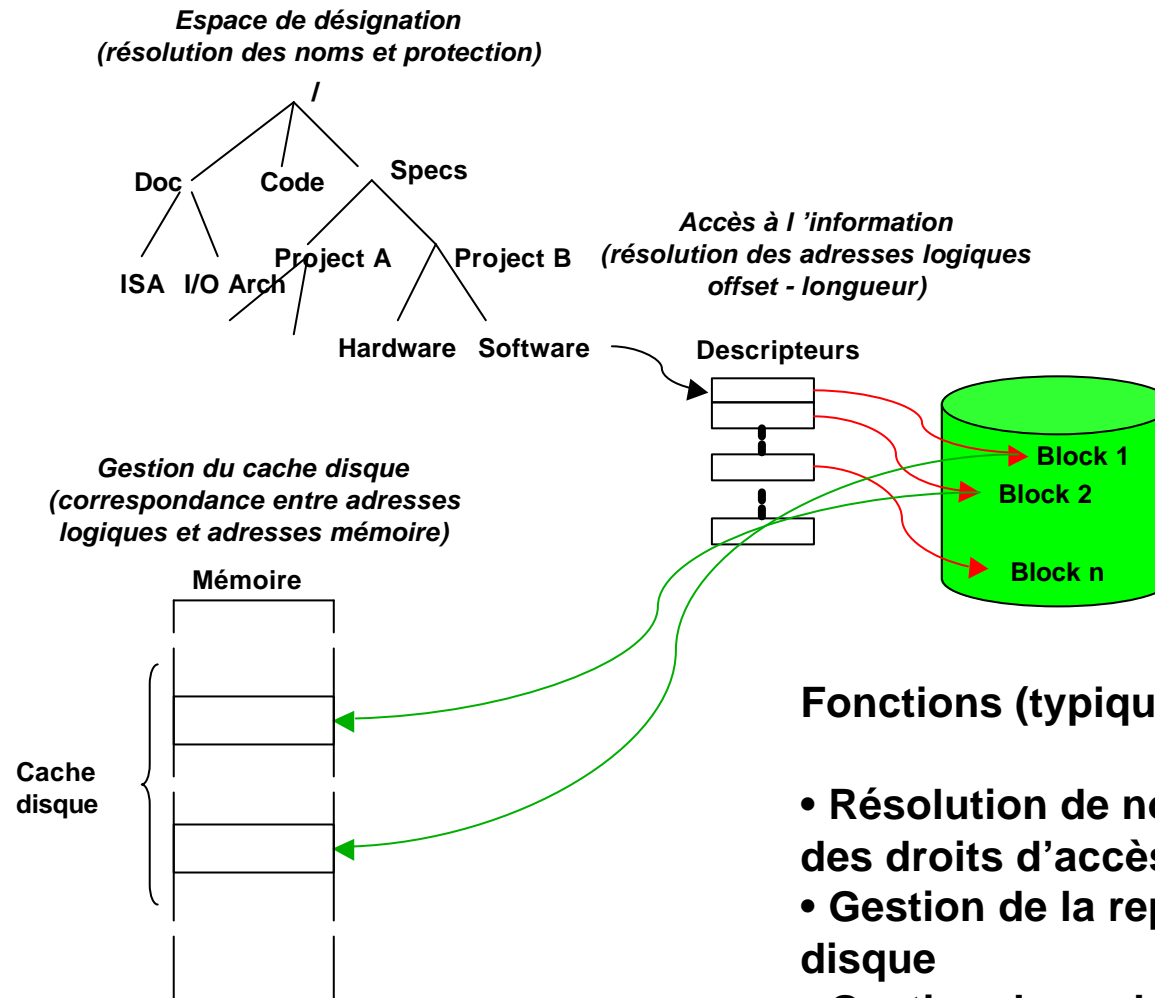
■ Structure simplifiée d'un disque



■ Temps d'accès (e.g. en lecture) :

- Positionnement du bras
- Délai rotationnel (1/2 tour en moyenne)
- Temps de transfert dans le tampon
- Temps de transfert vers le système demandeur (e.g. via un bus SCSI)

■ Fonctionnalité d'un système de fichiers



Fonctions (typiques) :

- Résolution de noms et gestion des droits d'accès
- Gestion de la représentation sur disque
- Gestion du cache disque en mémoire

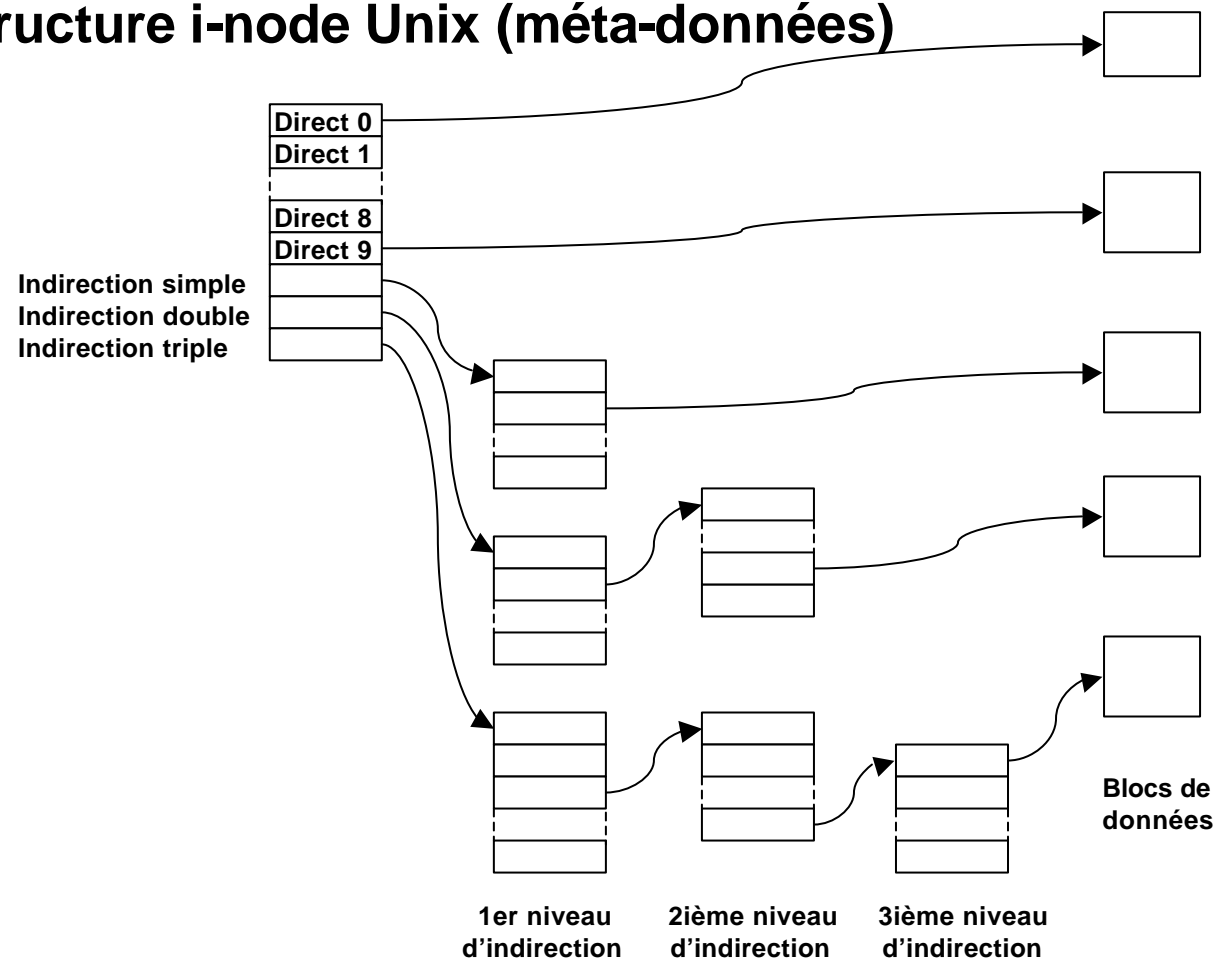
- **L'usage veut que l'on désigne sous le vocable « File System » (système de fichiers) deux notions différentes :**
 - Le logiciel en charge de la gestion de fichiers proprement dite, qu'il conviendrait donc de désigner sous le terme de « système de gestion de fichiers » ou « File Management System »
 - Un ensemble de fichiers
- Dans le cadre de cette présentation, on en cherchera pas à aller contre cet usage très répandu mais à donner suffisamment de contexte pour lever toute ambiguïté
- **Les systèmes de gestion de fichiers tels que ceux d'UNIX ou de Windows intègrent une méthode d'accès élémentaire (n octets à partir du p ième). Il en résulte une confusion entre système de gestion de fichiers (dont les fonctions ont été décrites précédemment) et les méthodes d'accès e.g. méthodes de base telles que celles d'Unix ou de Windows 2000, méthodes plus élaborées telles que le séquentiel indexé (C-ISAM, VSAM dans le monde propriétaire IBM)**

Systeme de fichiers(3)

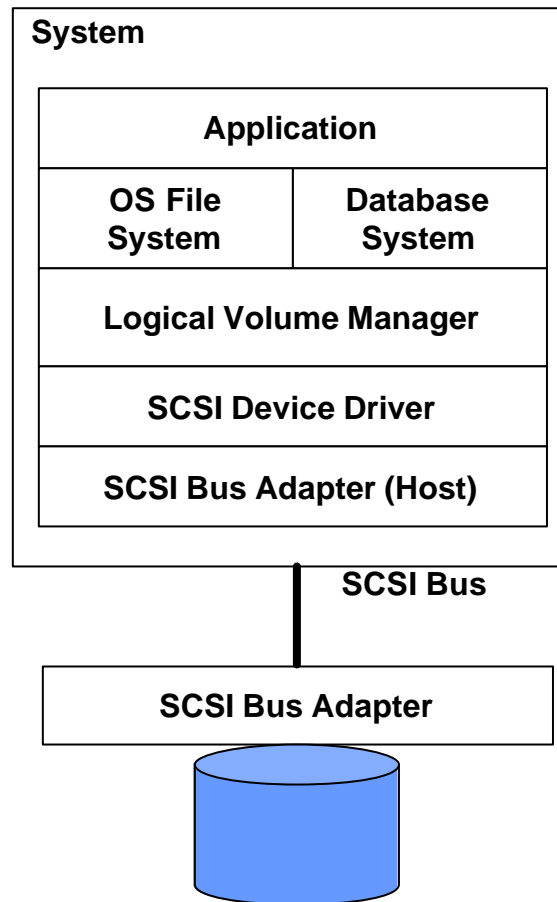
■ Structure g n rique d'un « File System » Unix



■ Structure i-node Unix (m ta-donn es)

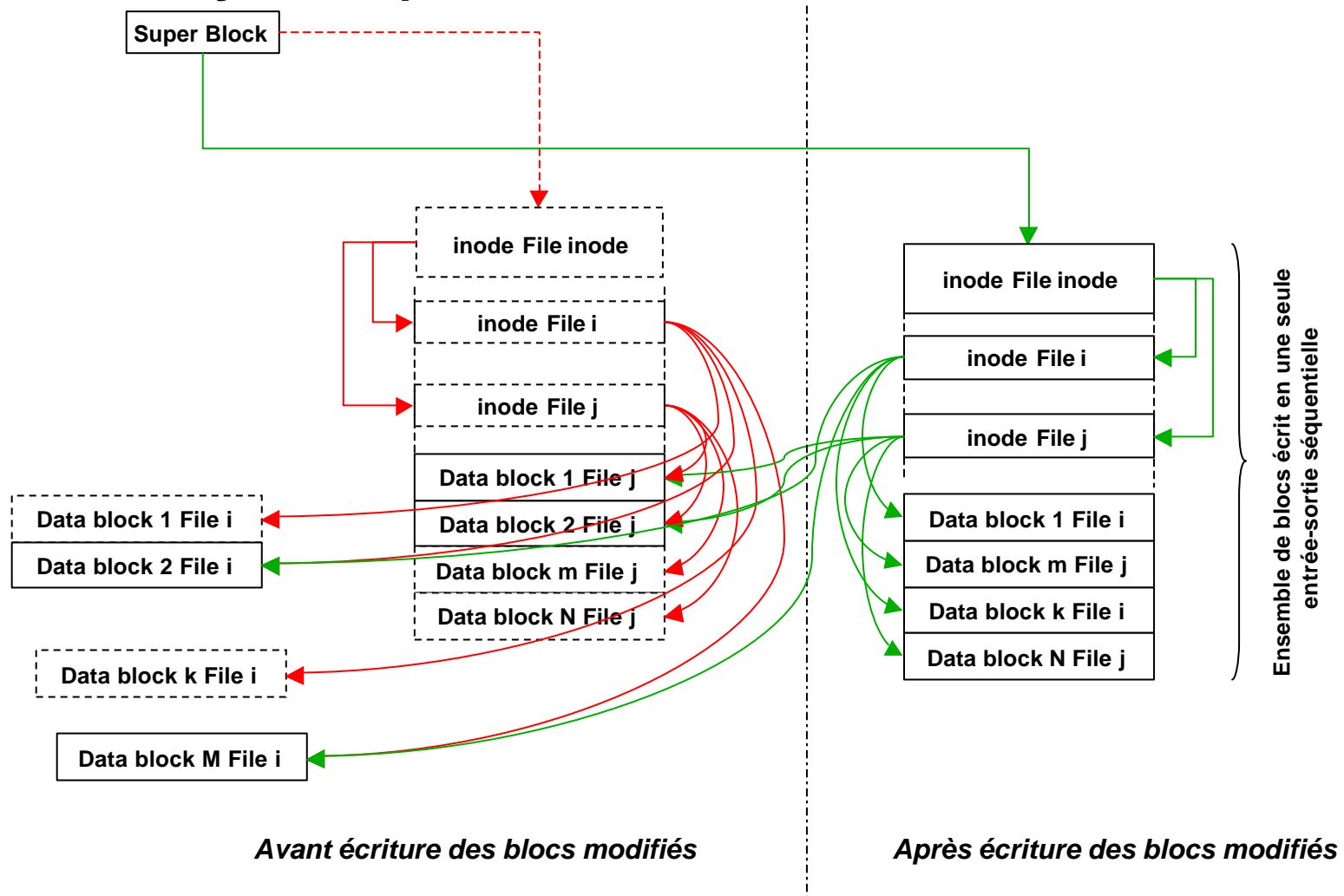


■ **Niveaux fonctionnels vis à vis de l'accès aux données (exemple) :**



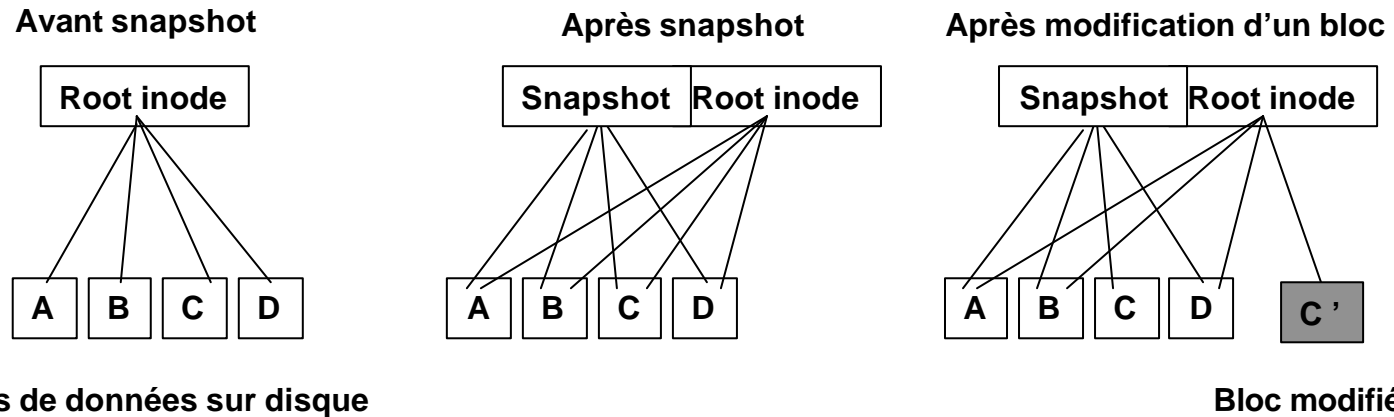
- **Systeme de fichiers journalisé :**
 - Avec un système de gestion de fichiers « classique », un incident en cours de fonctionnement provoque, lors de la ré-initialisation du système, une recherche séquentielle sur l'ensemble des disques pour vérifier la cohérence des structures de données (méta-données) décrivant le système de fichiers (utilitaire *fsck* sur Unix)
 - La vérification des structures de données prend un temps important (plusieurs dizaines de minutes)
 - Pour diminuer les temps de reprise, des systèmes de gestion de fichiers dit « journalisés » ont vu le jour. Dans ces systèmes, toutes les opérations entraînant des modifications des méta-données sont considérées comme des transaction et sont donc susceptibles d'être annulées ou rejouées

■ Systeme de fichiers optimisé (Log Structured File System)



Systeme de fichiers(7)

■ Mécanisme de copie instantanée (snapshot)



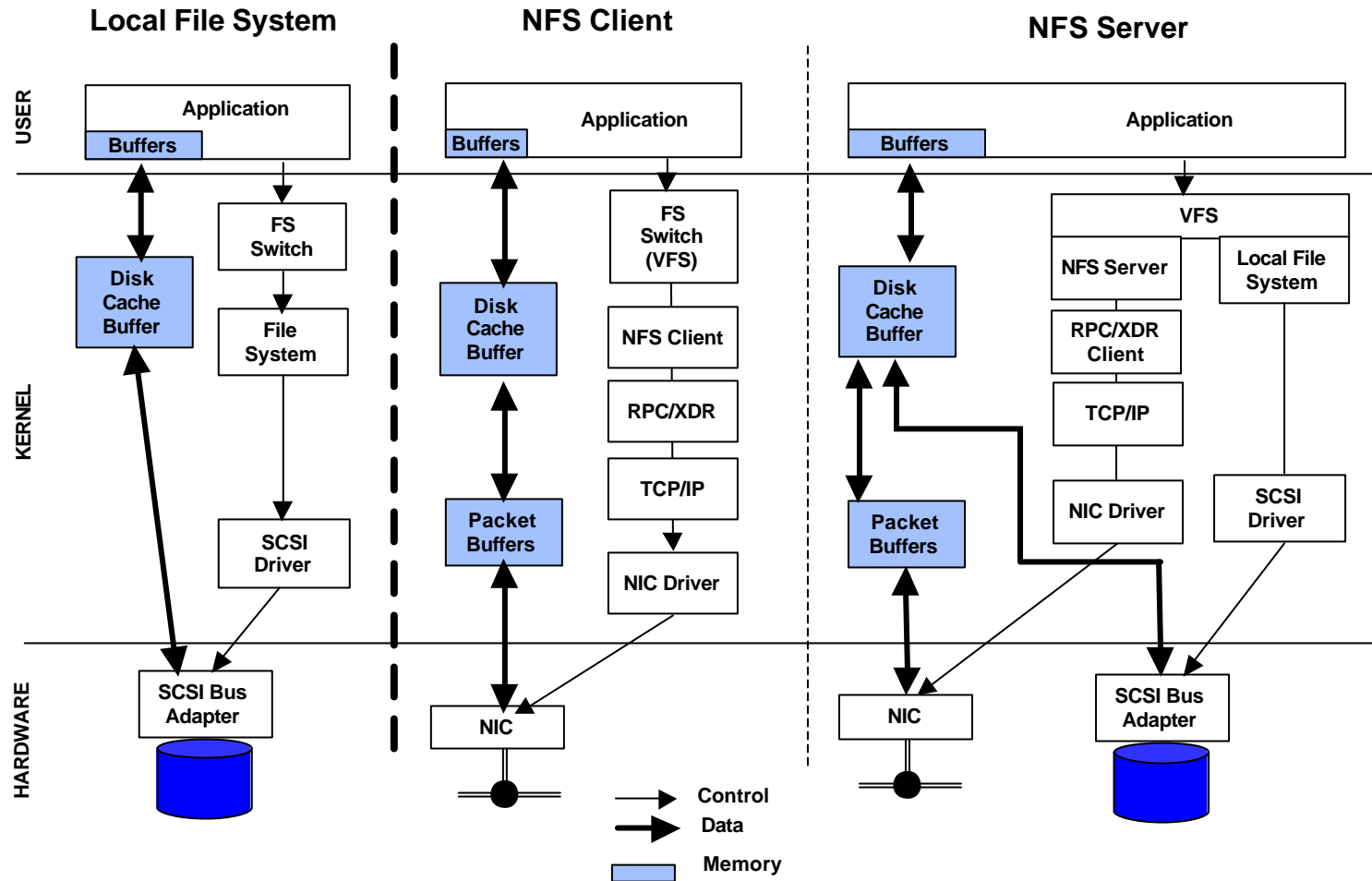
■ Avantages et inconvénients des Log Structured File Systems

	Log Structured File System	File System « classique »
Avantages	<ul style="list-style-type: none"> ◆ Amélioration des performances en écriture (minimisation des mouvements des bras des disques) ◆ Facilite les opérations de sauvegarde (snapshots, incrémentale, différentielle,..) 	<ul style="list-style-type: none"> ◆ Implémentations largement répandues ◆ Technologie éprouvée
Inconvénients	<ul style="list-style-type: none"> ◆ Nécessite une implémentation spécifique et plus complexe ◆ Paramétrage des opérations de récupération de l'espace 	<ul style="list-style-type: none"> ◆ Performances en écriture ◆ Opérations de sauvegarde plus complexes

Accès aux fichiers distants NFS, CIFS et DAFS

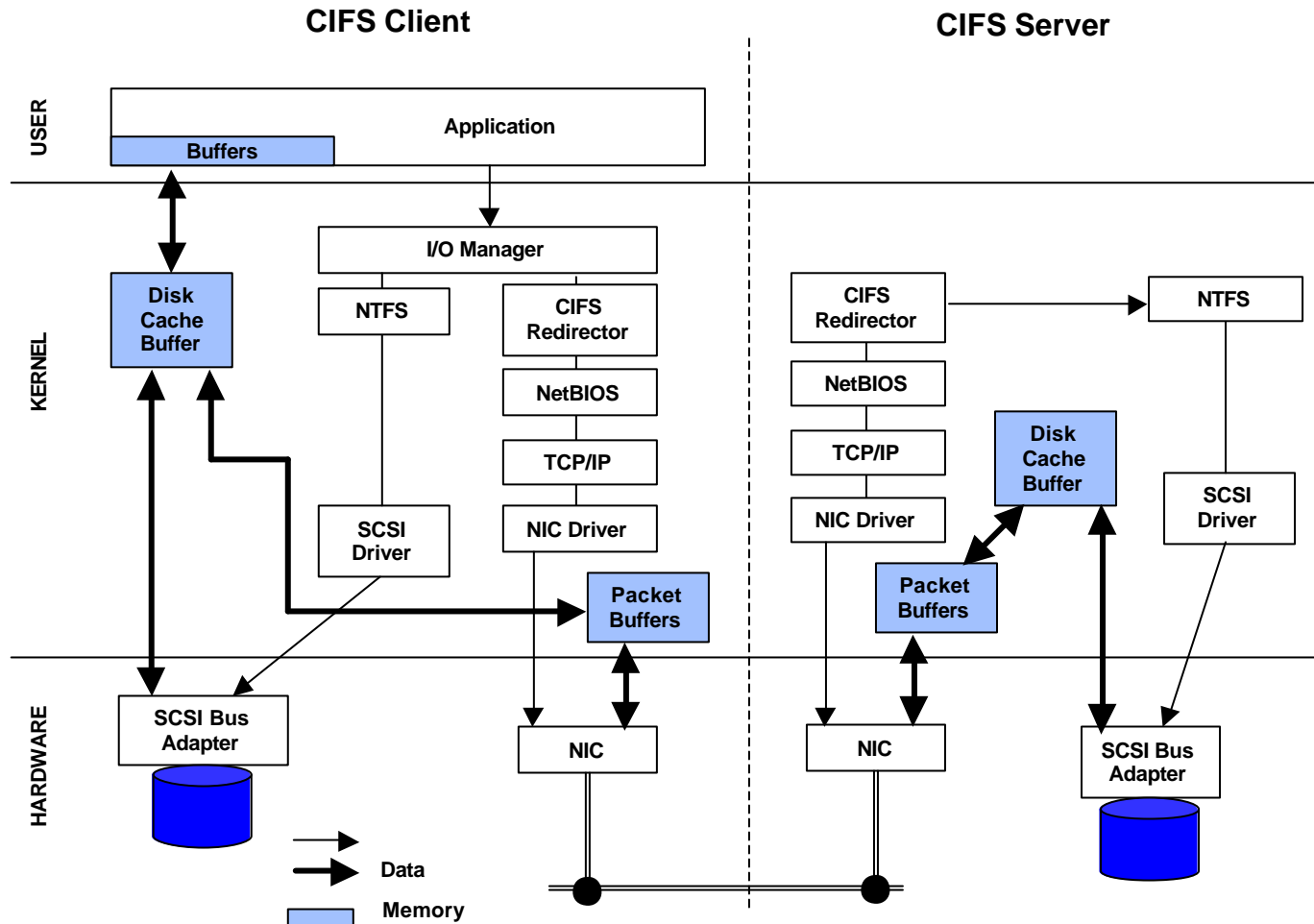
NFS - Network File System

- Couches logicielles mises en jeu dans l'accès aux fichiers locaux et distants avec NFS



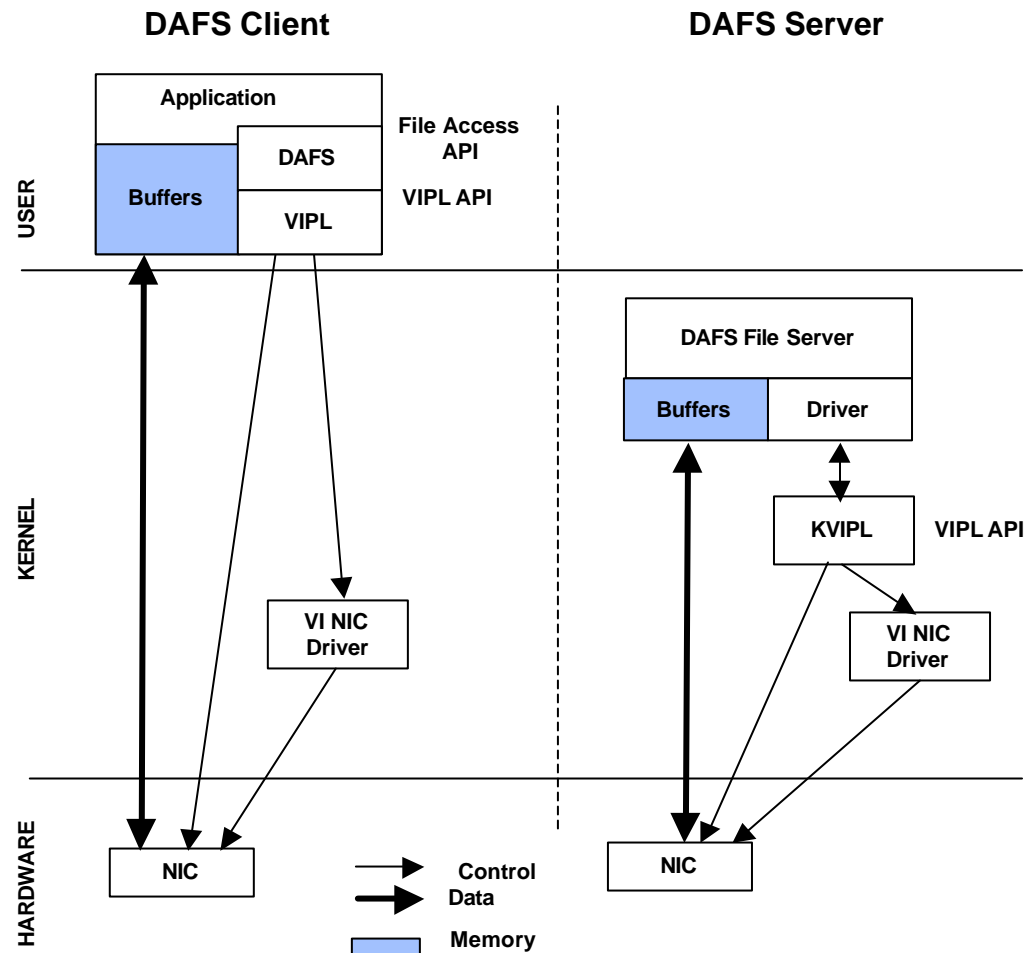
CIFS - Common Internet File System

■ Architecture CIFS - Windows



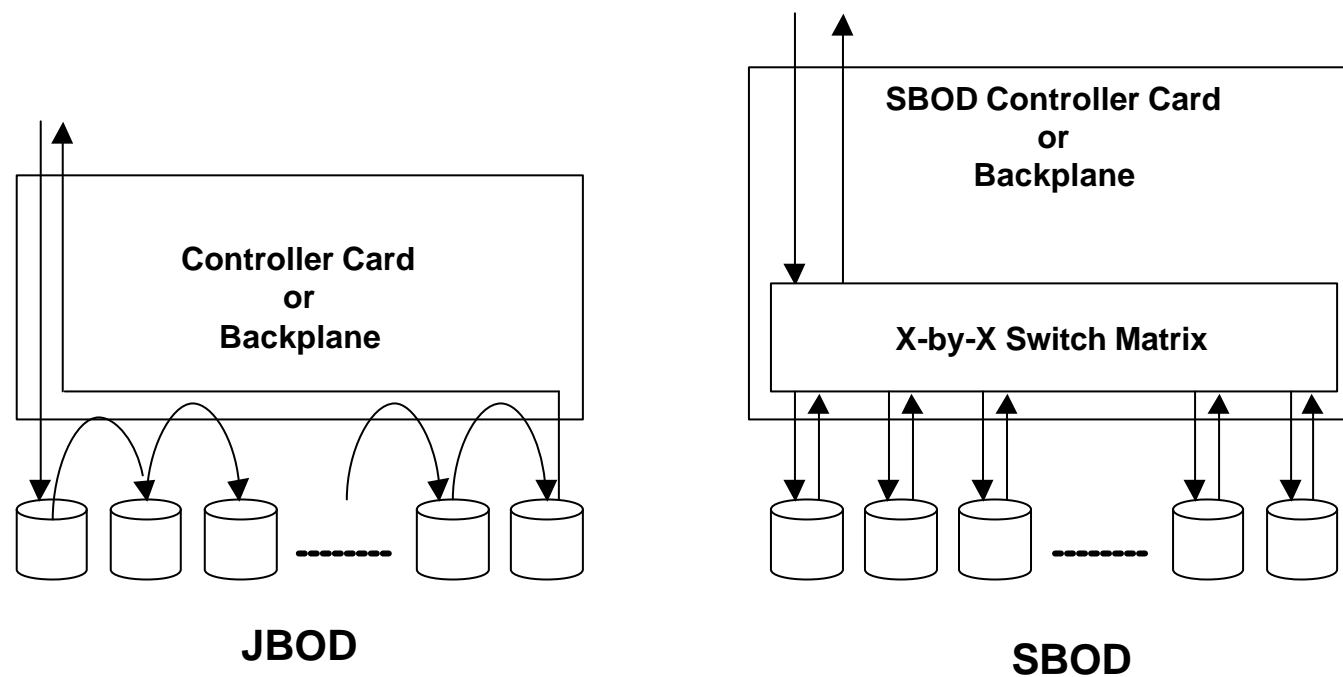
DAFS - Direct Access File System

- Couches logicielles mise en jeu dans l'accès aux fichiers avec DAFS



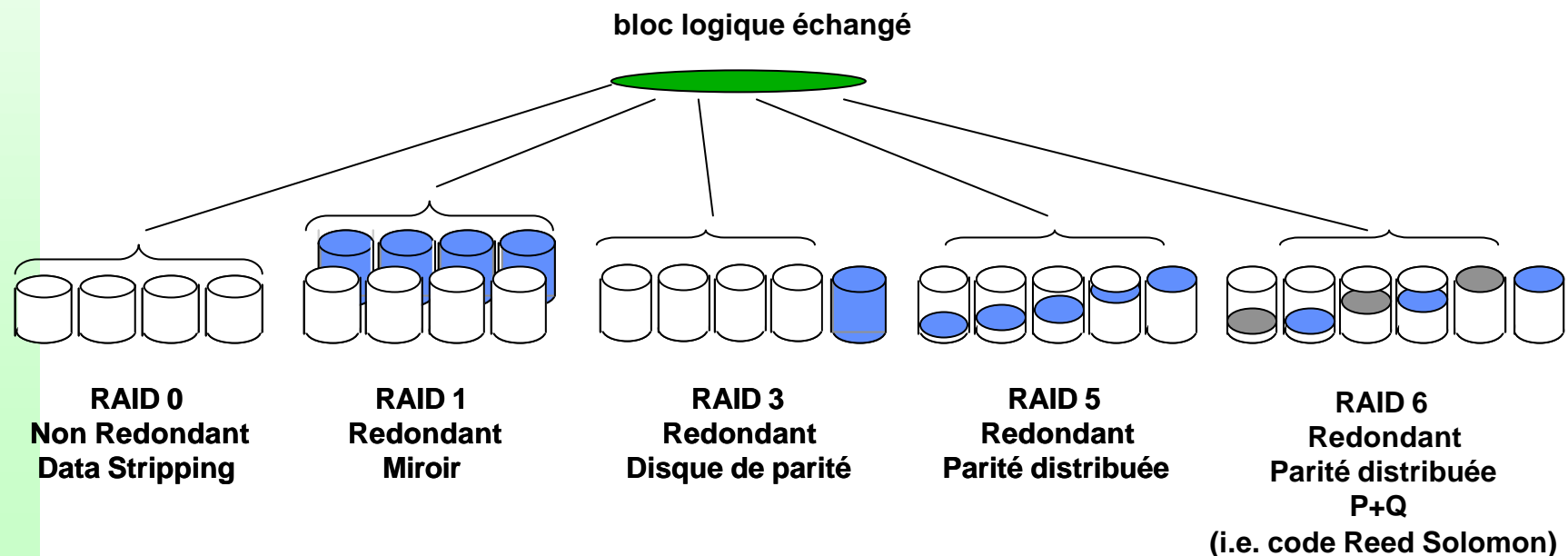
Organisation des disques JBOD, SBOD et RAID

- **JBOD = Just a Bunch Of Disks**
 - Objectif : capacité et coût réduit
- **SBOD = Switched Bunch Of Disks**
 - Objectif : capacité, disponibilité et performance



Technologie RAID - Tableaux de disques

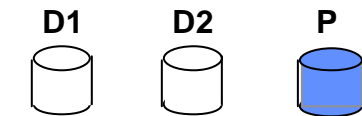
- **RAID : Redundant Array of Independent (Inexpensive) Disks**
- **La technologie RAID a été formalisée par des chercheurs de l'Université de Berkeley [PAT88]**
- **Principe: groupement de petits disques pour constituer un ensemble de grande capacité, à grande performance et à haute disponibilité :**
 - Répartition des données sur plusieurs disques et transferts en parallèle
 - Redondance économique (utilisation de disques de parité)
- **On présente ici les niveaux de RAID les plus fréquemment utilisés (parmi les 7 niveaux identifiés de RAID 0 à RAID 6). Le choix entre les différents niveaux de RAID dépend de l'utilisation (voir page suivante)**



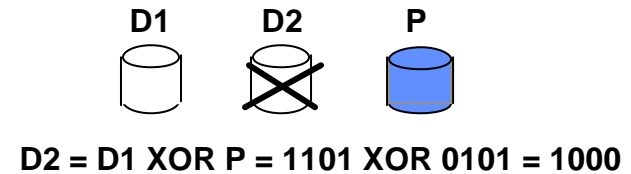
■ Redondance fondée sur la technique du ou exclusif (XOR)

	0	1
0	0	1
1	1	0

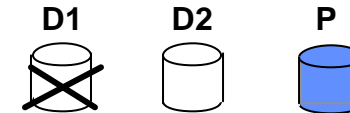
Rappel : définition XOR



$$P = 1101 \text{ XOR } 1000 = 0101$$



$$D2 = D1 \text{ XOR } P = 1101 \text{ XOR } 0101 = 1000$$



$$D1 = D2 \text{ XOR } P = 1000 \text{ XOR } 0101 = 1101$$

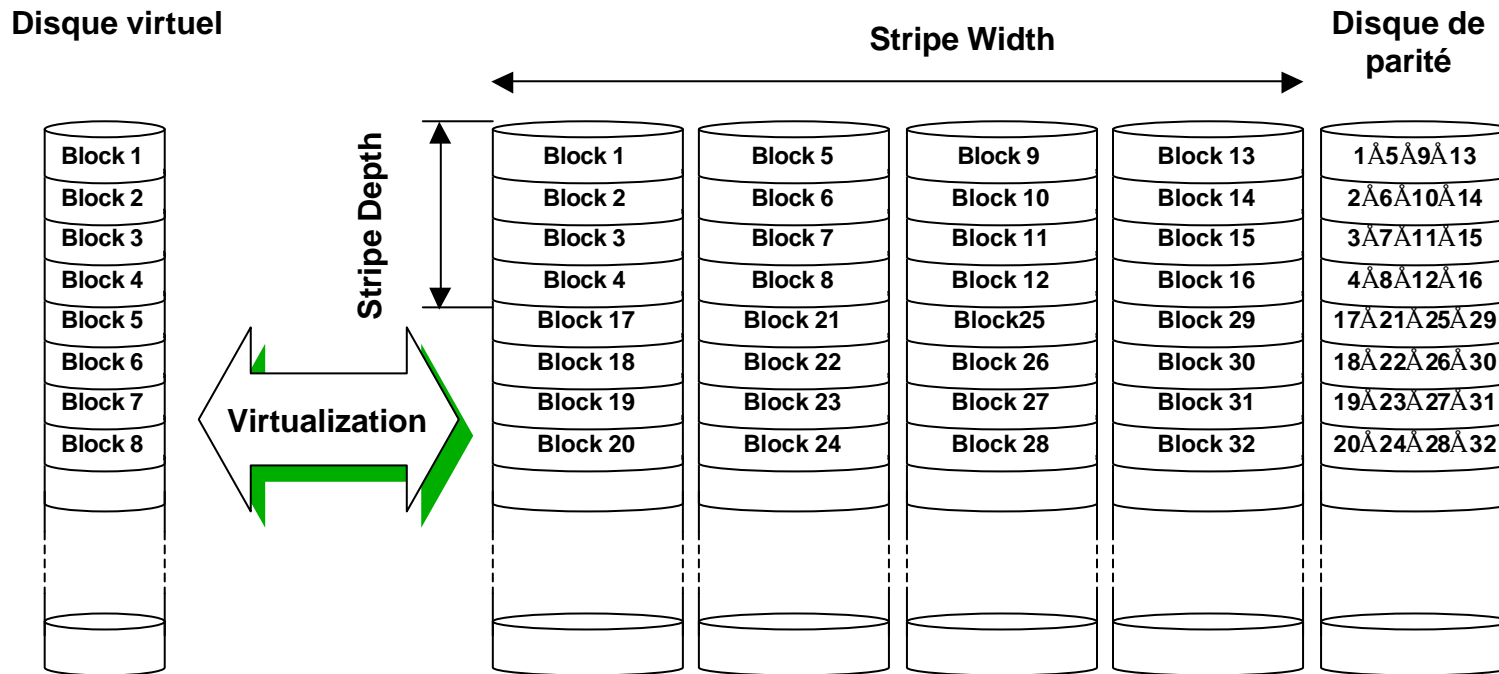
■ Cas d'utilisation :

- RAID 0: performance sans redondance
- RAID 1: performance et redondance coûteuse (2 x disques)
- RAID 3: redondance économique (1 disque de parité pour n disques de données) et performance pour les grands transferts de données
- RAID 5: redondance économique (1 disque de parité pour n disques de données) et performance pour les petits transferts de données
- RAID 6: mêmes caractéristiques que RAID 5 mais capacité à résister à la défaillance de deux disques.

- **Application « naïve » du concept RAID = disque de très grande capacité donc introduction d'un niveau de virtualisation**
- **Avantages de la virtualisation :**
 - ◆ Fournir aux système(s) d'exploitation la vision d'un stockage sous forme de disques virtuels dont les tailles et les niveaux de fonctionnalité RAID peuvent être choisies en fonction des besoins
 - ◆ Support, avec une vision commune de disques virtuels, plusieurs unités RAID dont les disques ont des caractéristiques différentes
 - ◆ Tirer le meilleur profit des ressources physiques installées e.g. 5 disques pouvant être accédés en parallèle auxquels on ajoute 5 nouveaux disques pouvant eux aussi être accédés en parallèle : réorganisation du disque virtuel sur 10 disques accédés en parallèle

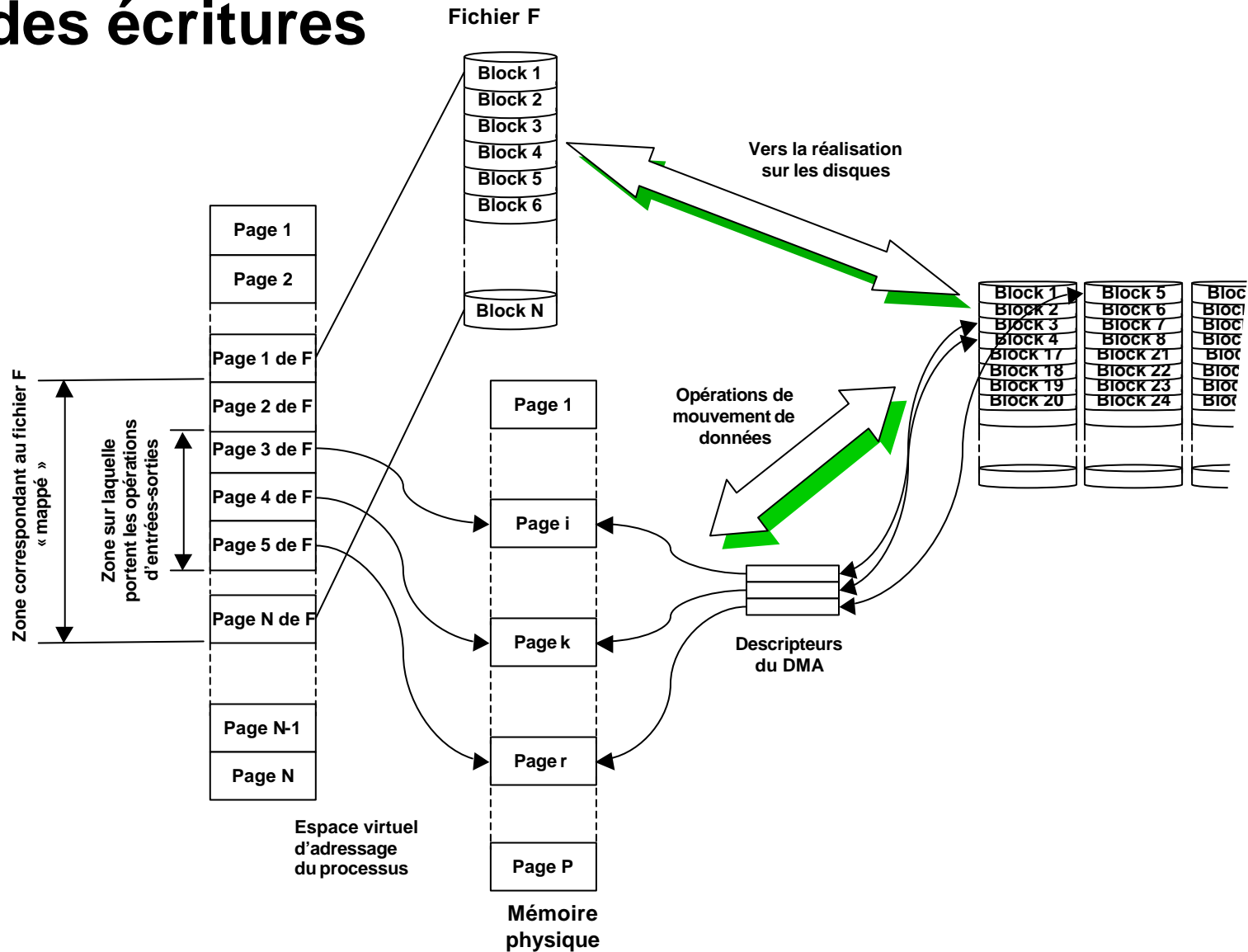
Virtualisation du stockage(2)

■ Implémentation du RAID avec virtualisation des disques



Note : La largeur du stripe (4 ici) et sa profondeur (4 aussi dans cet exemple) sont indépendantes

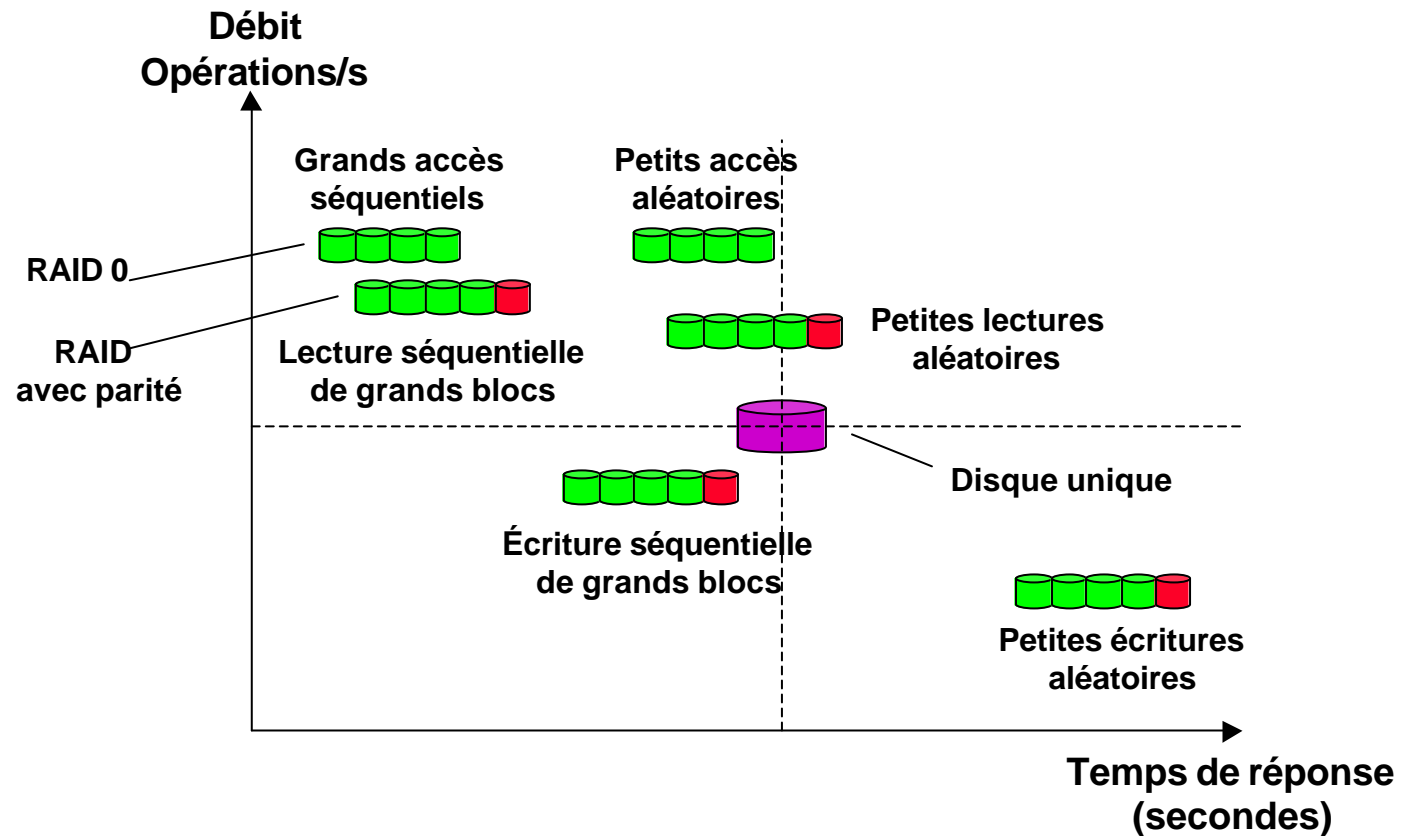
■ Regroupement des lectures et éclatement des écritures



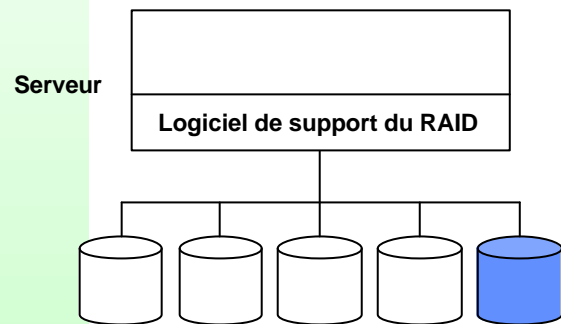
Comparaison des niveaux de RAID usuels

Nom	Coût du stockage	Disponibilité relative des données	Rapidité des grandes lectures séquentielles	Rapidité des grandes écritures séquentielles	Débit en lectures aléatoires	Débit en écritures aléatoires
Data striping	>1	Inférieure à celle d'une organisation conventionnelle	Elevée - Dépend du nombre de disques en parallèle	Elevée - Dépend du nombre de disques en parallèle	Elevé	Elevé
Mirroring (d'ordre M, M=2 le plus souvent)	x M	>RAID 3, RAID 5 <RAID 6	Jusqu'à M fois un disque unique	Inférieur à disque unique	Jusqu'à M fois un disque unique	Inférieur à disque unique
Stripped mirrors (miroir d'ordre M, M=2 le plus souvent)	M x N	>RAID 3, RAID 5 <RAID 6	Jusqu'à M fois RAID 0 équivalent	Peut être supérieure à celle du disque unique en fonction de N	Jusqu'à M fois RAID 0 équivalent	Inférieur à RAID 0 équivalent
Disque de parité	N + 1	>> disque conventionnel	Elevée - Dépend du nombre de disques en parallèle (<RAID 0)	Elevée - Dépend du nombre de disques en parallèle et nécessite le calcul de la parité (<RAID 0)	Elevé	Nécessite le calcul et la mise à jour de l'information de parité
Parité «spiralee»	N + 1	>> disque conventionnel ~ RAID 3	< RAID 0 en raison du contrôle de parité	< RAID 0	Elevé > RAID 3	>> RAID 3
Double parité «spiralee»	N + 2	Supérieure à tous les autres types	Légèrement > RAID 5	< RAID 5 (2 informations de parité)	Légèrement > RAID 5	< RAID 5 (2 informations de parité)

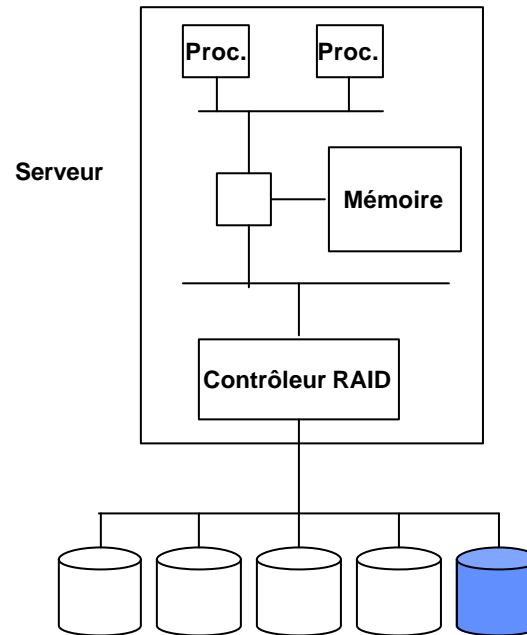
■ Positionnement en performance



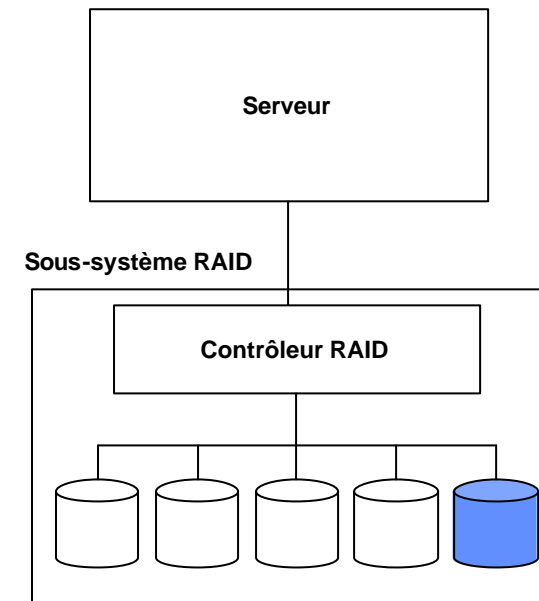
■ Options d'architecture pour le support de la fonction RAID



RAID supporté par le serveur



RAID supporté par un contrôleur au sein du serveur

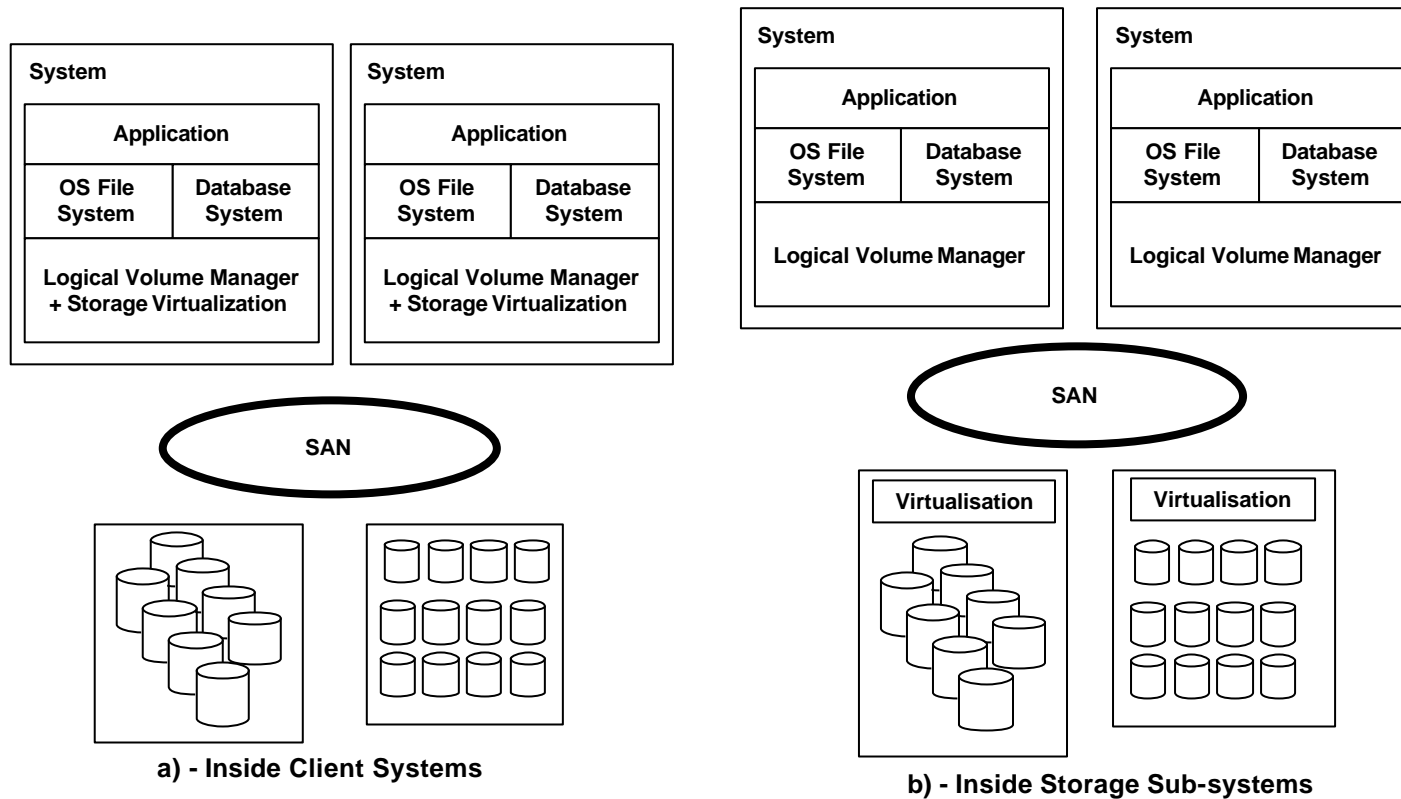


RAID supporté au sein d'un sous-système indépendant

Comparaison des options d'architecture

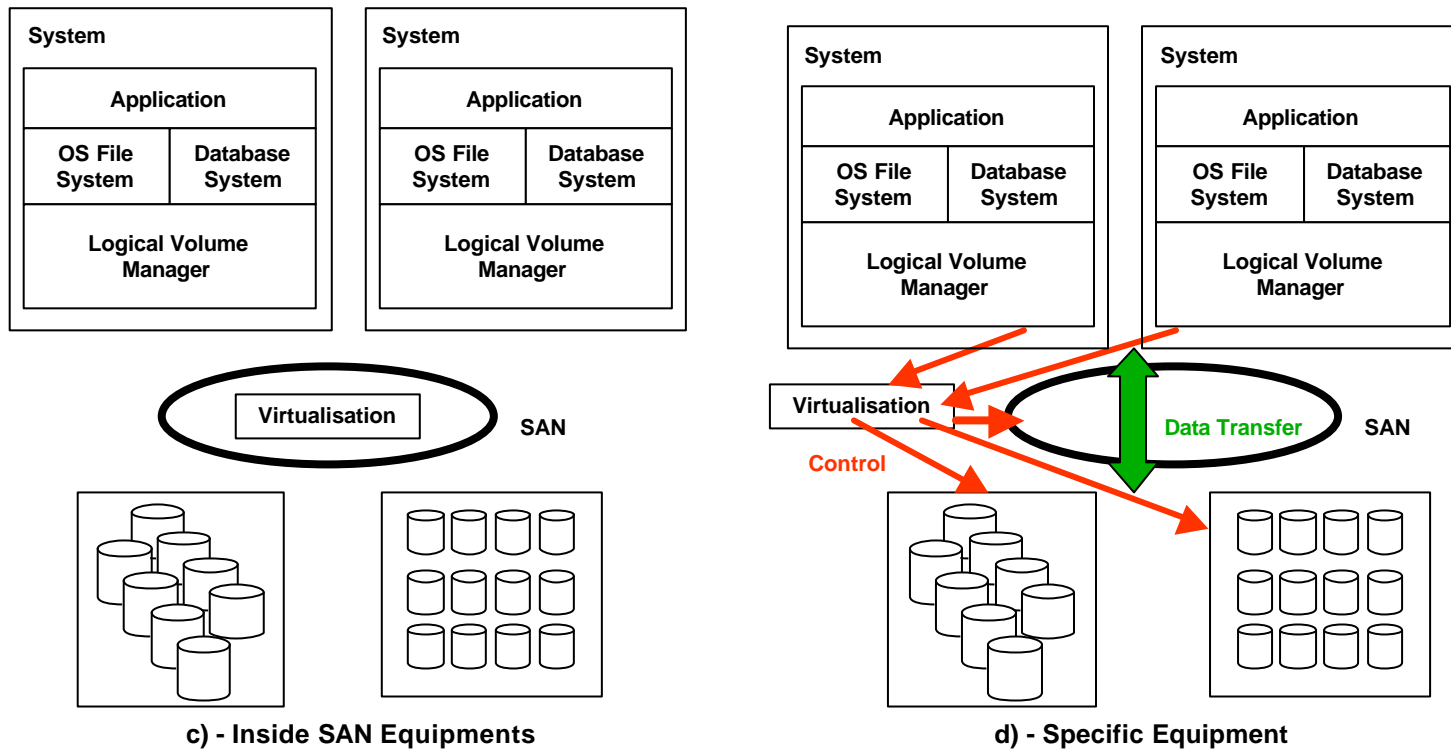
RAID supporté par le serveur	RAID supporté par contrôleur dans serveur	RAID supporté au sein d'un sous-système spécialisé
<ul style="list-style-type: none"> . Faible coût . Capacité de raccordement élevée (i.e. celle du serveur) . Scalabilité (l'accroissement de la performance du serveur profite au support du RAID) . Bonne disponibilité (faible nombre de composants mis en jeu) 	<ul style="list-style-type: none"> . Coût modéré . Bons temps d'exécution et bon débit (matériel spécialisé) 	<ul style="list-style-type: none"> . Capacité de raccordement généralement importante (celle du sous-système) et possibilité de supporter plusieurs sous-systèmes . Débit élevé (matériel spécialisé) . Bon temps de réponse pour les opérations d'écriture si un cache d'écriture est supporté . Disponibilité élevée (doublement des contrôleurs internes, chemins d'accès multiples) . Indépendance entre le moyen de connexion des systèmes hôtes (e.g. Fibre Channel) et des disques (e.g. SCSI)
<ul style="list-style-type: none"> . La performance du serveur est impactée par le support de la fonction RAID . La disponibilité des données impose de doubler les chemins d'accès au disques (récupération des disques suite à la défaillance du serveur) 	<ul style="list-style-type: none"> . Nombre de disques supporté limité aux capacités de raccordement des contrôleurs dans le système . La disponibilité des données impose de doubler les chemins d'accès au disques (récupération des disques suite à la défaillance du contrôleur ou du serveur) 	<ul style="list-style-type: none"> . Architecture matérielle spécifique (cache sécurisé redondant) . Coût élevé . Temps de réponse supérieur à celui de la solution "contrôleur" en raison de la liaison serveur/sous-système

Options d'architecture pour le support de la virtualisation



Support de la virtualisation(2)

Options d'architecture pour le support de la virtualisation(2)

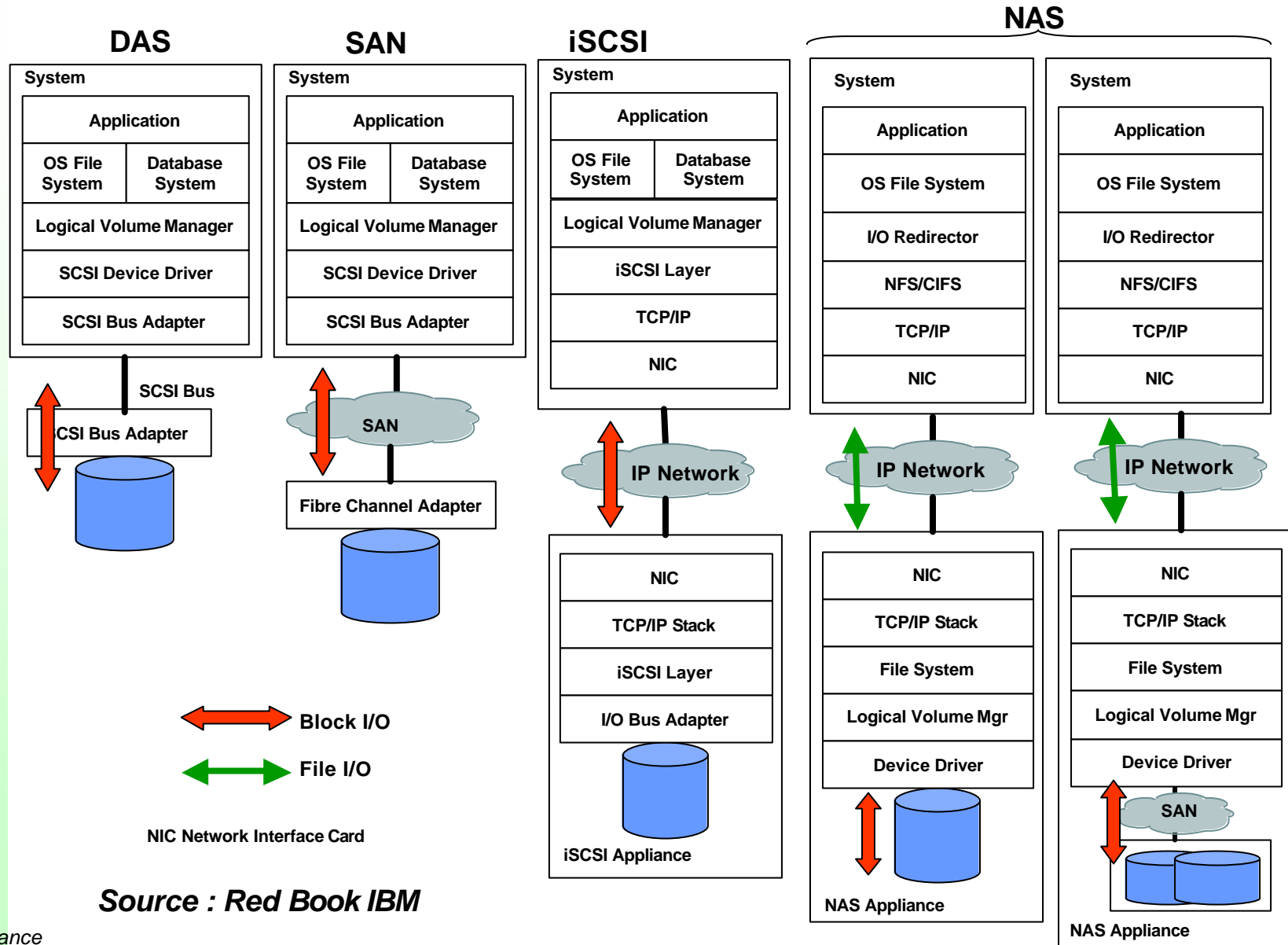


■ Comparaison des options

	Systèmes clients	Sous-systèmes de stockage	Equipements du SAN	Equipement spécifique
A V A N T A G E S	<p>Virtualisation fondée sur des techniques éprouvées</p> <p>Intégration étroite avec les Files Systems et les SGBDs</p>	<p>Permet de supporter l'hétérogénéité du stockage (technologie et fournisseurs)</p>	<p>Possibilité de connecter des clients de natures diverses</p>	<p>Contrôle centralisé</p> <p>Haute performance du fait de la séparation du contrôle et des mouvements de données</p> <p>Support de clients hétérogènes</p>
I N C O N V E N I E N T S	<p>La visibilité globale du stockage impose de disposer de techniques de clusterisation</p> <p>Complexité d'administration</p>	<p>Plusieurs points d'administration</p> <p>Solution spécifique de chaque fournisseur</p> <p>La visibilité globale du stockage impose le recours à la clusterisation</p> <p>Qualification de la solution</p> <p>Coût des différents sous-systèmes</p>	<p>La visibilité globale du stockage impose de disposer de techniques de clusterisation</p> <p>Nécessité de techniques de clusterisation pour la continuité de service</p> <p>Nécessité des équipements susceptibles de supporter la fonction de virtualisation</p> <p>Interopérabilité entre les équipements de différents fournisseurs</p>	<p>Nécessite des pilotes spécifiques au niveau des clients</p> <p>Qualification de la solution en environnement hétérogène</p> <p>La haute disponibilité impose la redondance de l'équipement</p> <p>Complexité de la connectique</p>

Architectures de stockage DAS, SAN; NAS et iSCSI

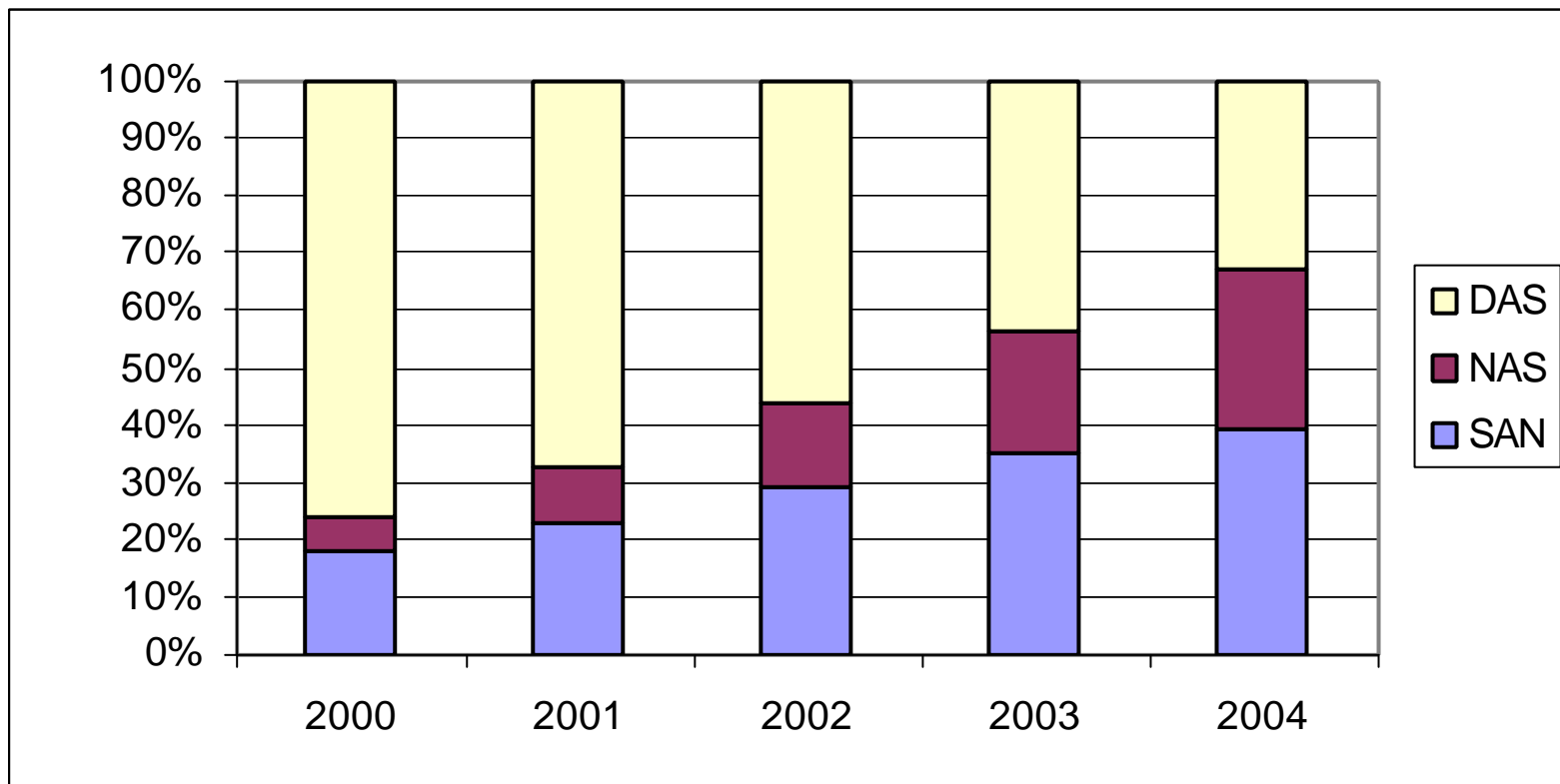
Illustration des architectures



Comparaison des architectures

	DAS	NAS	SAN	iSCSI
Type de liaison	<ul style="list-style-type: none"> ◆ SCSI ◆ FC-AL ◆ ... 	<ul style="list-style-type: none"> ◆ Fast Ethernet ◆ Fibre Channel 	<ul style="list-style-type: none"> ◆ Fibre Channel 	<ul style="list-style-type: none"> ◆ Internet
Connexion à distance	<ul style="list-style-type: none"> ◆ Typiquement non 	<ul style="list-style-type: none"> ◆ Oui 	<ul style="list-style-type: none"> ◆ Possible 	<ul style="list-style-type: none"> ◆ Oui
Type d'E/S	<ul style="list-style-type: none"> ◆ Niveau bloc 	<ul style="list-style-type: none"> ◆ Niveau fichier 	<ul style="list-style-type: none"> ◆ Niveau bloc 	<ul style="list-style-type: none"> ◆ Niveau bloc
Performance	<ul style="list-style-type: none"> ◆ Elevée 	<ul style="list-style-type: none"> ◆ Limitée par le réseau 	<ul style="list-style-type: none"> ◆ Elevée 	<ul style="list-style-type: none"> ◆ Limitée par le réseau
Partage des données	<ul style="list-style-type: none"> ◆ Implique NFS ou CIFS 	<ul style="list-style-type: none"> ◆ Natif 	<ul style="list-style-type: none"> ◆ Difficile (en 2002) /Futur 	<ul style="list-style-type: none"> ◆ Difficile (en 2002)
Réduction des coûts (mutualisation)	<ul style="list-style-type: none"> ◆ Non 	<ul style="list-style-type: none"> ◆ Oui 	<ul style="list-style-type: none"> ◆ Oui 	<ul style="list-style-type: none"> ◆ Oui
Séparation des investissements	<ul style="list-style-type: none"> ◆ Non 	<ul style="list-style-type: none"> ◆ Oui 	<ul style="list-style-type: none"> ◆ Oui 	<ul style="list-style-type: none"> ◆ Oui
Scalabilité	<ul style="list-style-type: none"> ◆ Non 	<ul style="list-style-type: none"> ◆ Oui 	<ul style="list-style-type: none"> ◆ Oui 	<ul style="list-style-type: none"> ◆ Dépend du support réseau
Disponibilité des données	<ul style="list-style-type: none"> ◆ Limitée 	<ul style="list-style-type: none"> ◆ Oui si redondance 	<ul style="list-style-type: none"> ◆ Oui si redondance 	<ul style="list-style-type: none"> ◆ Oui si redondance
Centralisation de la gestion et du support	<ul style="list-style-type: none"> ◆ Typiquement non 	<ul style="list-style-type: none"> ◆ Oui 	<ul style="list-style-type: none"> ◆ Oui 	<ul style="list-style-type: none"> ◆ Oui
Gestion	<ul style="list-style-type: none"> ◆ Traditionnelle 	<ul style="list-style-type: none"> ◆ Type SNMP 	<ul style="list-style-type: none"> ◆ Difficile (en 2002) 	<ul style="list-style-type: none"> ◆ Difficile (en 2002)
LAN Free Backup	<ul style="list-style-type: none"> ◆ Non 	<ul style="list-style-type: none"> ◆ Dépend du serveur NAS 	<ul style="list-style-type: none"> ◆ Oui 	<ul style="list-style-type: none"> ◆ Dépend du serveur iSCSI
Server Free Backup	<ul style="list-style-type: none"> ◆ Non 	<ul style="list-style-type: none"> ◆ Dépend du serveur NAS 	<ul style="list-style-type: none"> ◆ Oui 	<ul style="list-style-type: none"> ◆ Dépend du serveur iSCSI
Sécurité	<ul style="list-style-type: none"> ◆ Par le serveur 	<ul style="list-style-type: none"> ◆ Par les serveurs et le réseau 	<ul style="list-style-type: none"> ◆ Par les serveurs et le réseau de stockage 	<ul style="list-style-type: none"> ◆ Par les serveurs et le réseau
Installation	<ul style="list-style-type: none"> ◆ Spécifique au serveur 	<ul style="list-style-type: none"> ◆ Aisée 	<ul style="list-style-type: none"> ◆ Difficile (en 2002) 	<ul style="list-style-type: none"> ◆ Difficile (en 2002)

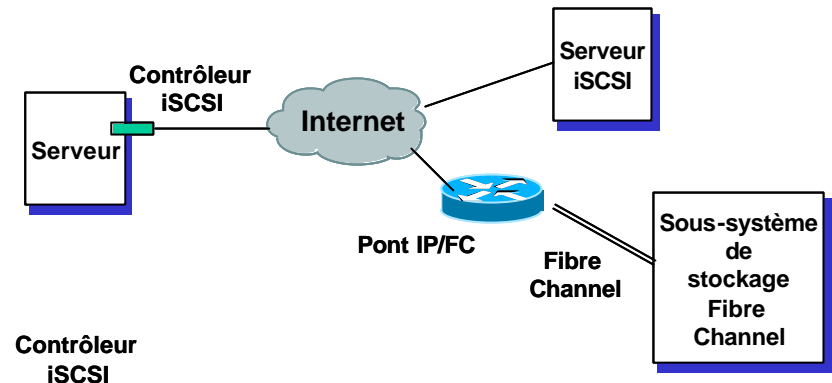
■ Évolution des parts de marché des différentes solutions (Source IDC)



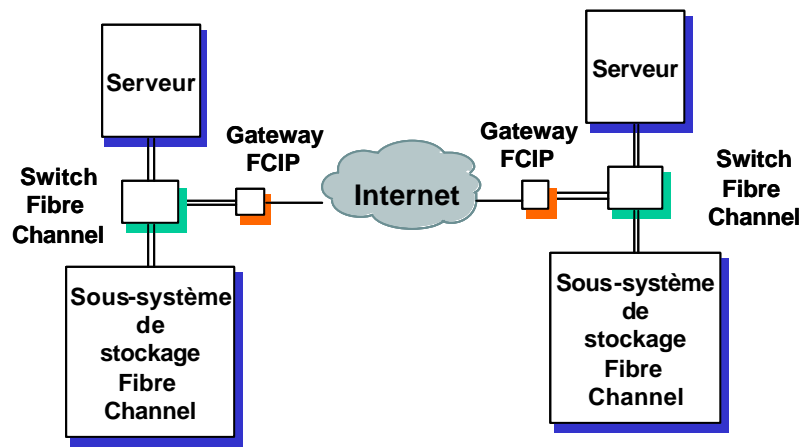
Marché mondial 2001 = \$39B, 2004 = \$53B

Intégration Internet et Fibre Channel

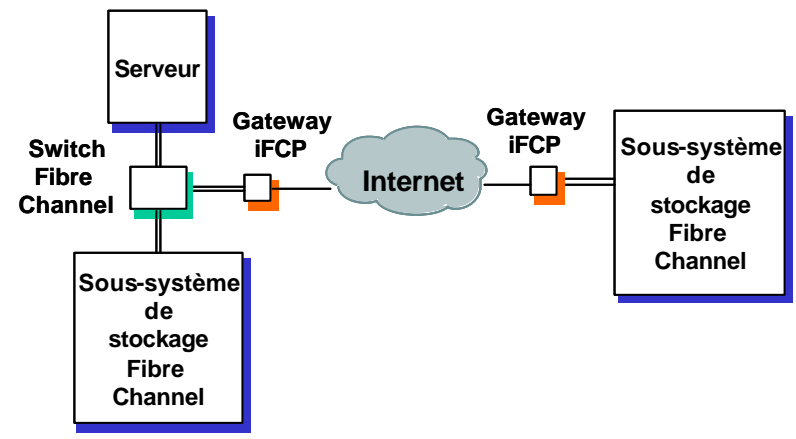
■ Illustration des intégrations possibles



a) – Réseau SAN sur Internet

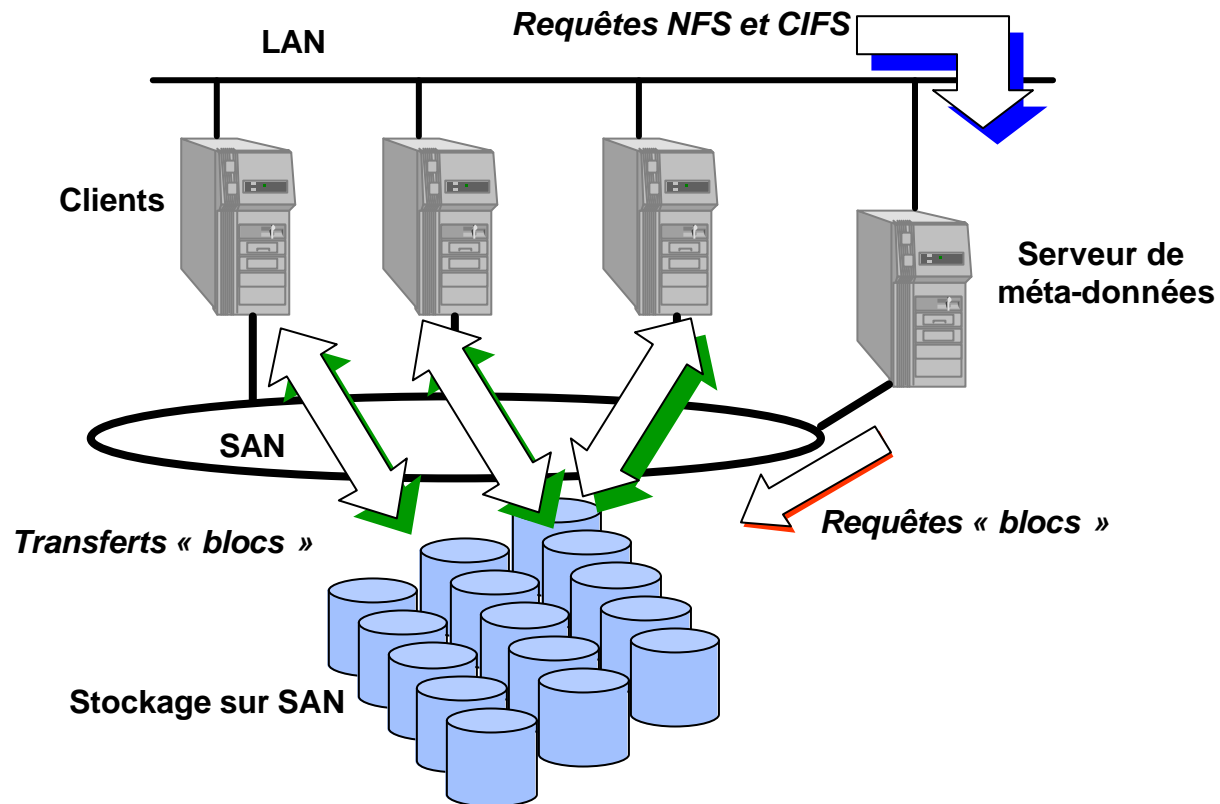


b) – Interconnexion de SAN via Internet

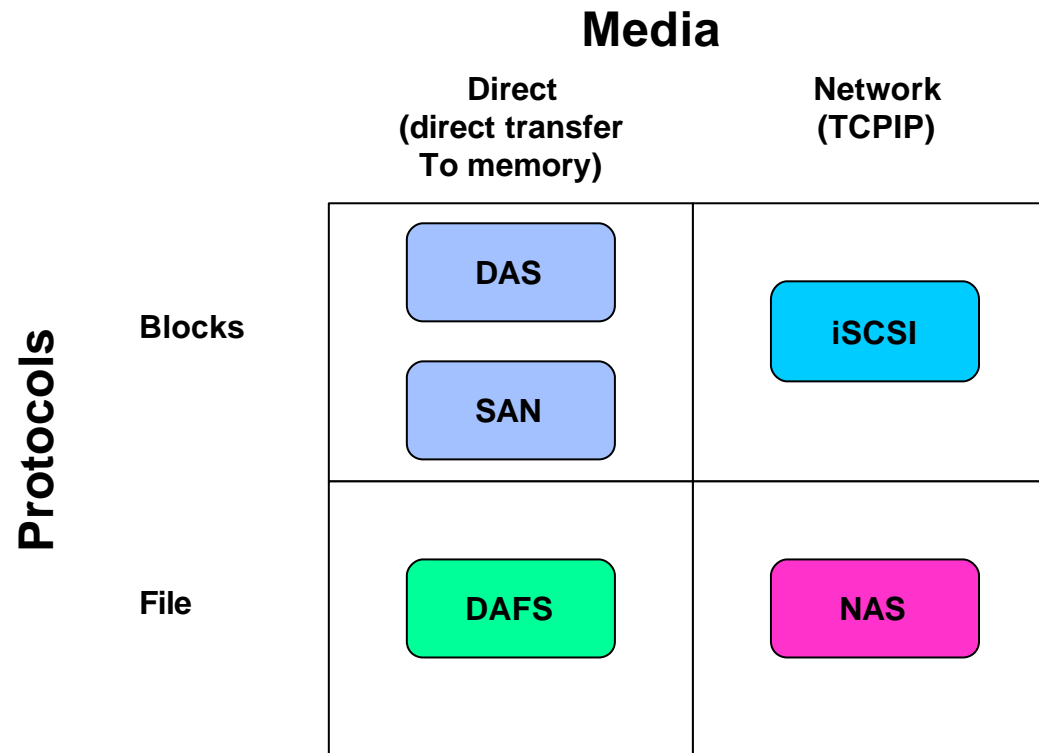


c) – Fibre Channel sur Internet

■ Intégration SAN et NAS dans une architecture commune

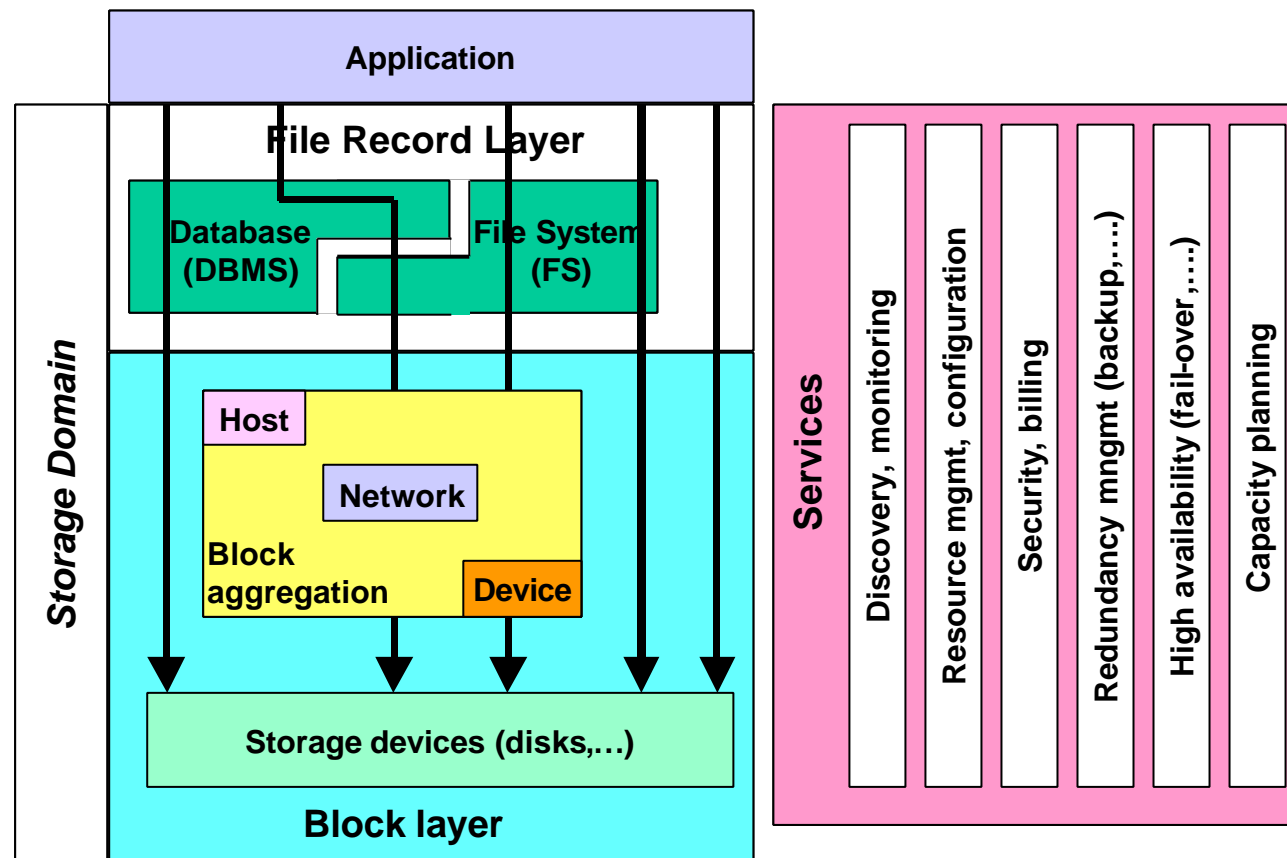


■ Synthèse des options d'architecture de stockage



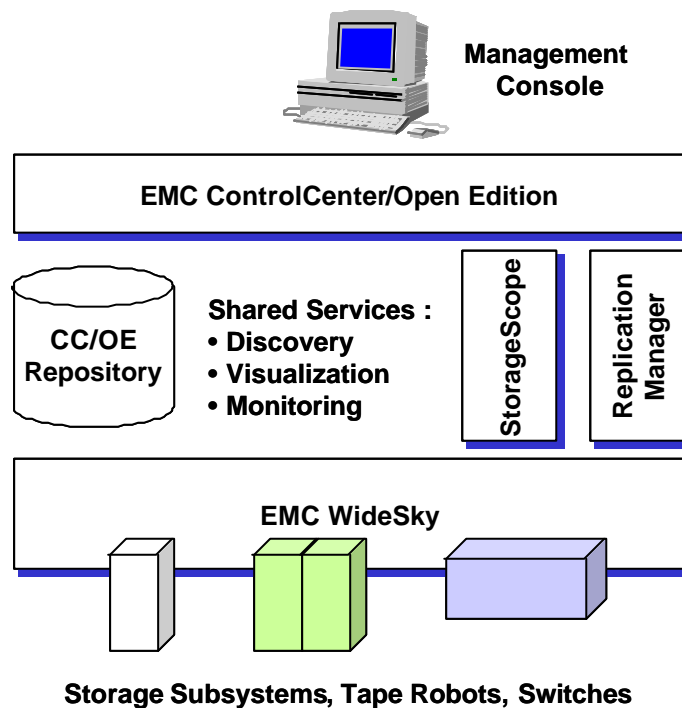
Modèle d'architecture SNIA

- **Modèle d'architecture SNIA (Storage Network Industry Association) : “ Shared Storage Model - A Framework for Describing Storage Architectures ”**



Management du stockage

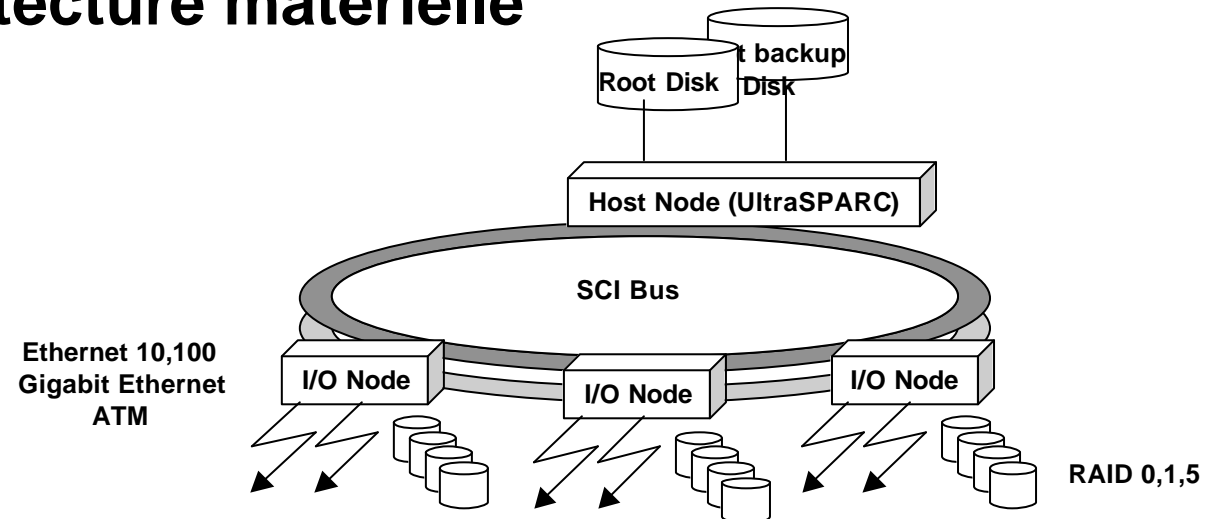
■ Exemple de l'architecture de WideSky (EMC)



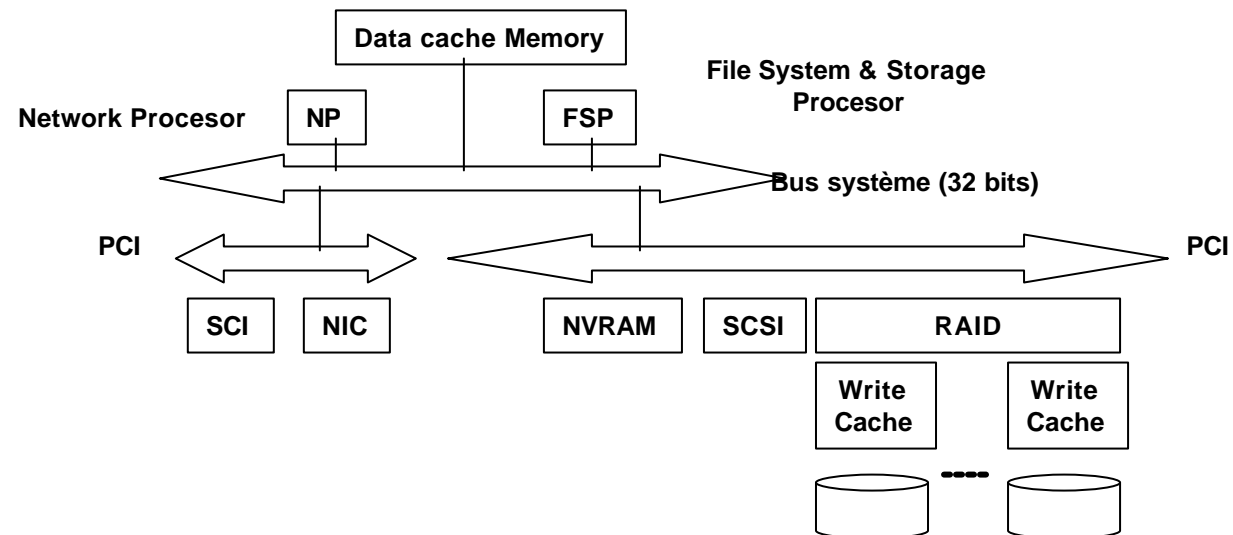
- Vision homogène des différents équipements ;
- ControlCenter/Open Edition Repository : base de données contenant les informations sur l'ensemble des éléments ;
- Services partagés par les composants de l'offre :
 - découverte de la configuration,
 - visualisation,
 - surveillance du fonctionnement,...
- StorageScope : reporting de l'utilisation des ressources de stockage ;
- Replication Manager : outil de gestion des réplications ;
- ControlCenter/Open Edition : outil de gestion centralisé des ressources

Exemples de solutions de stockage

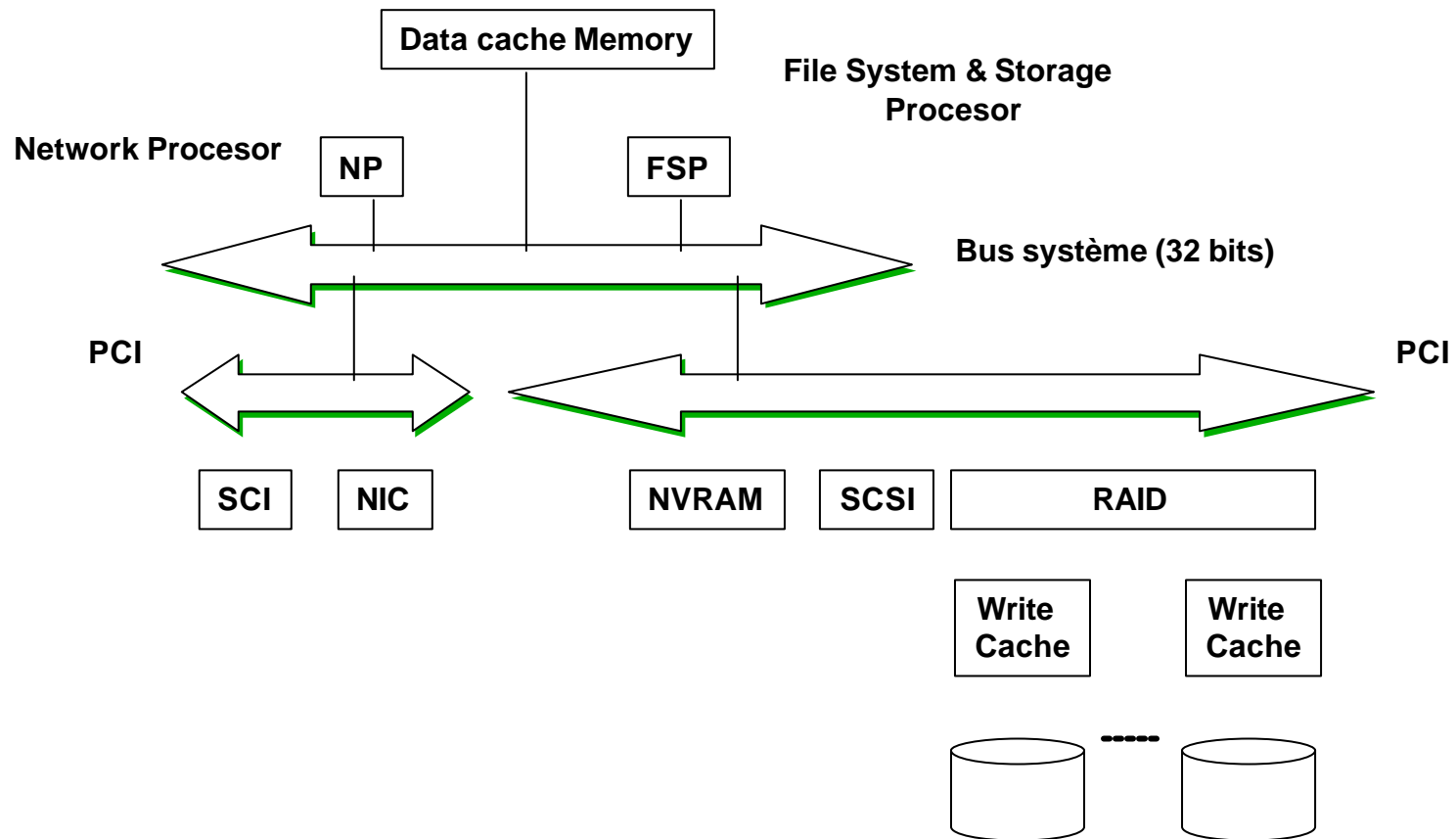
■ Architecture matérielle



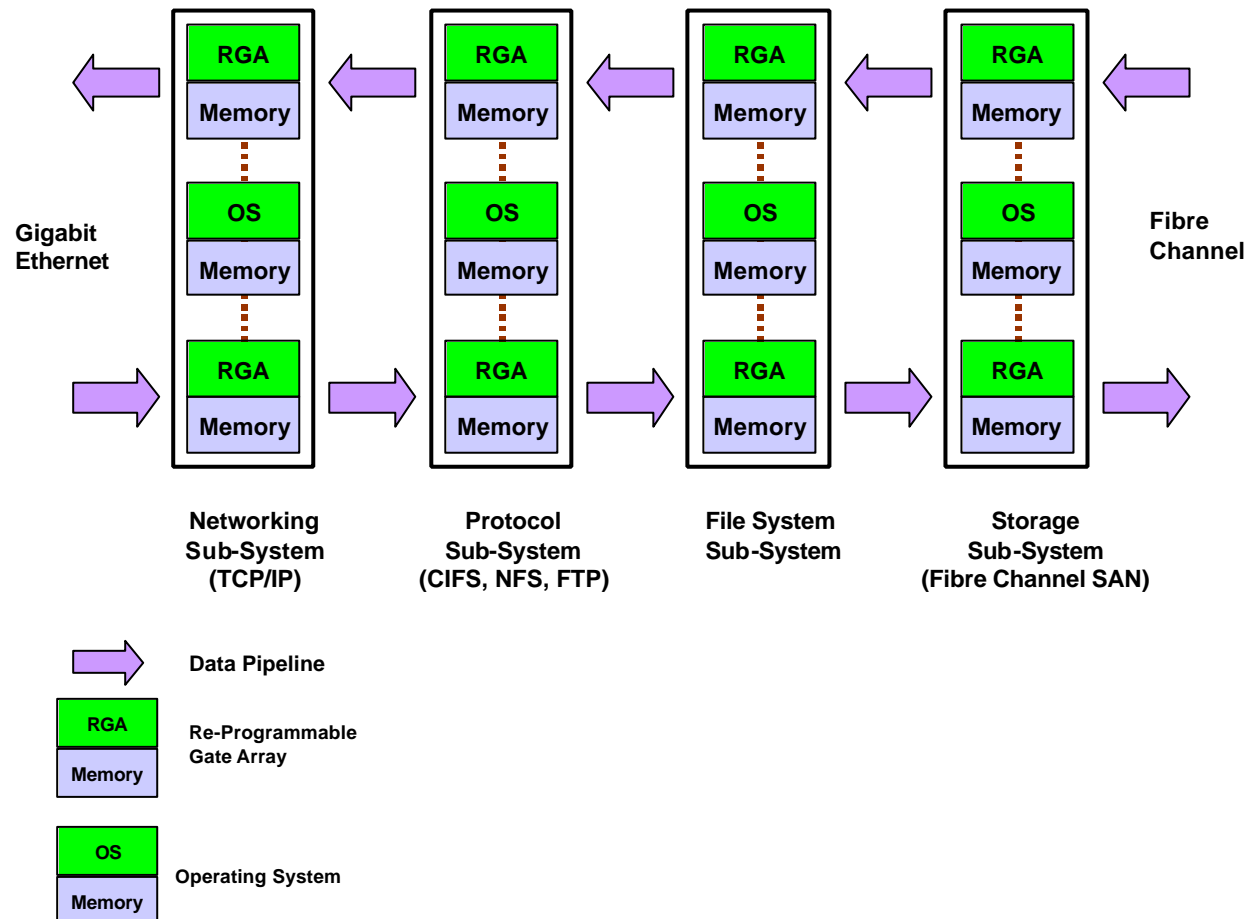
■ Architecture I/O Node



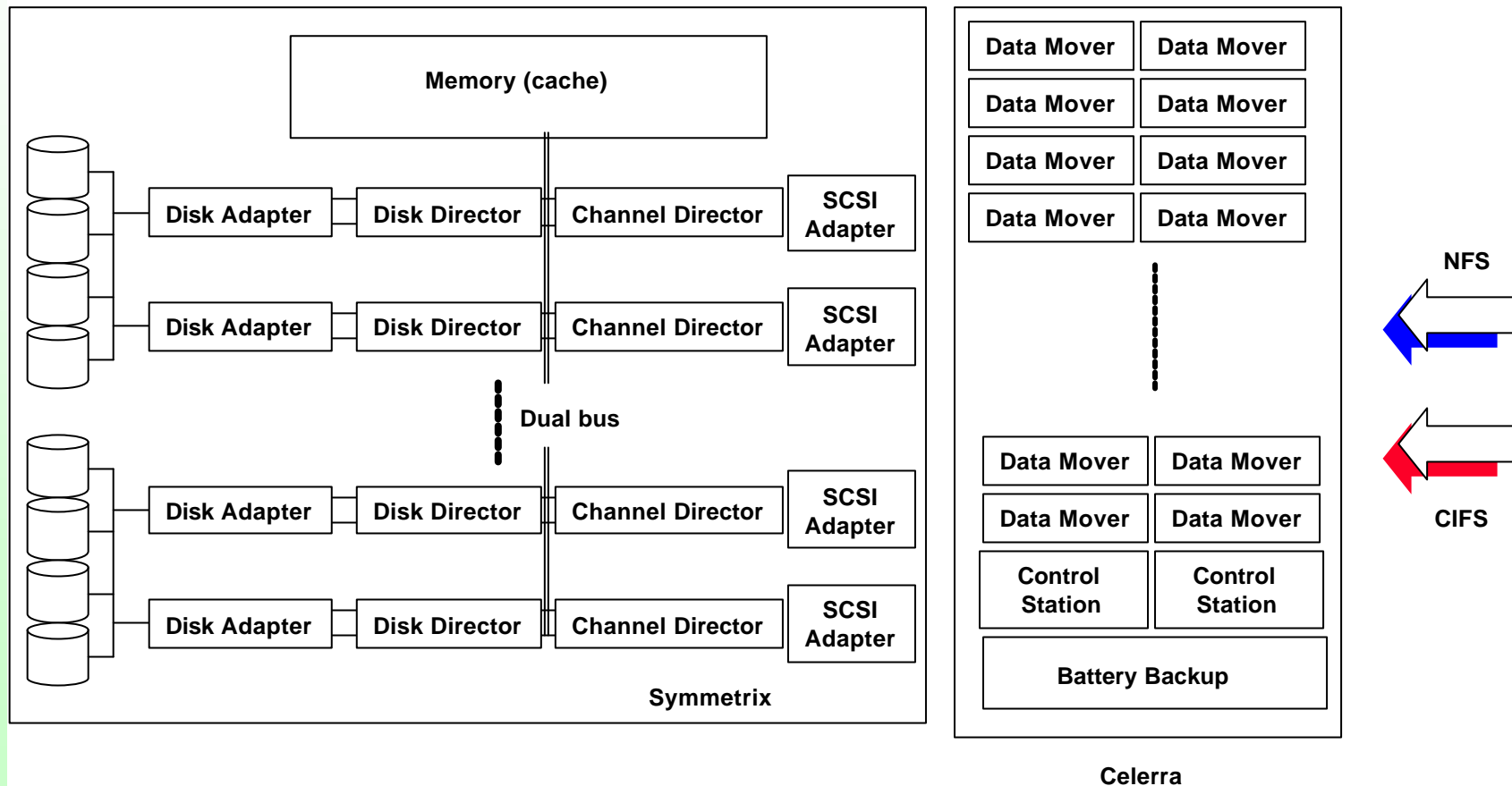
■ Architecture du nœud d'entrées-sorties



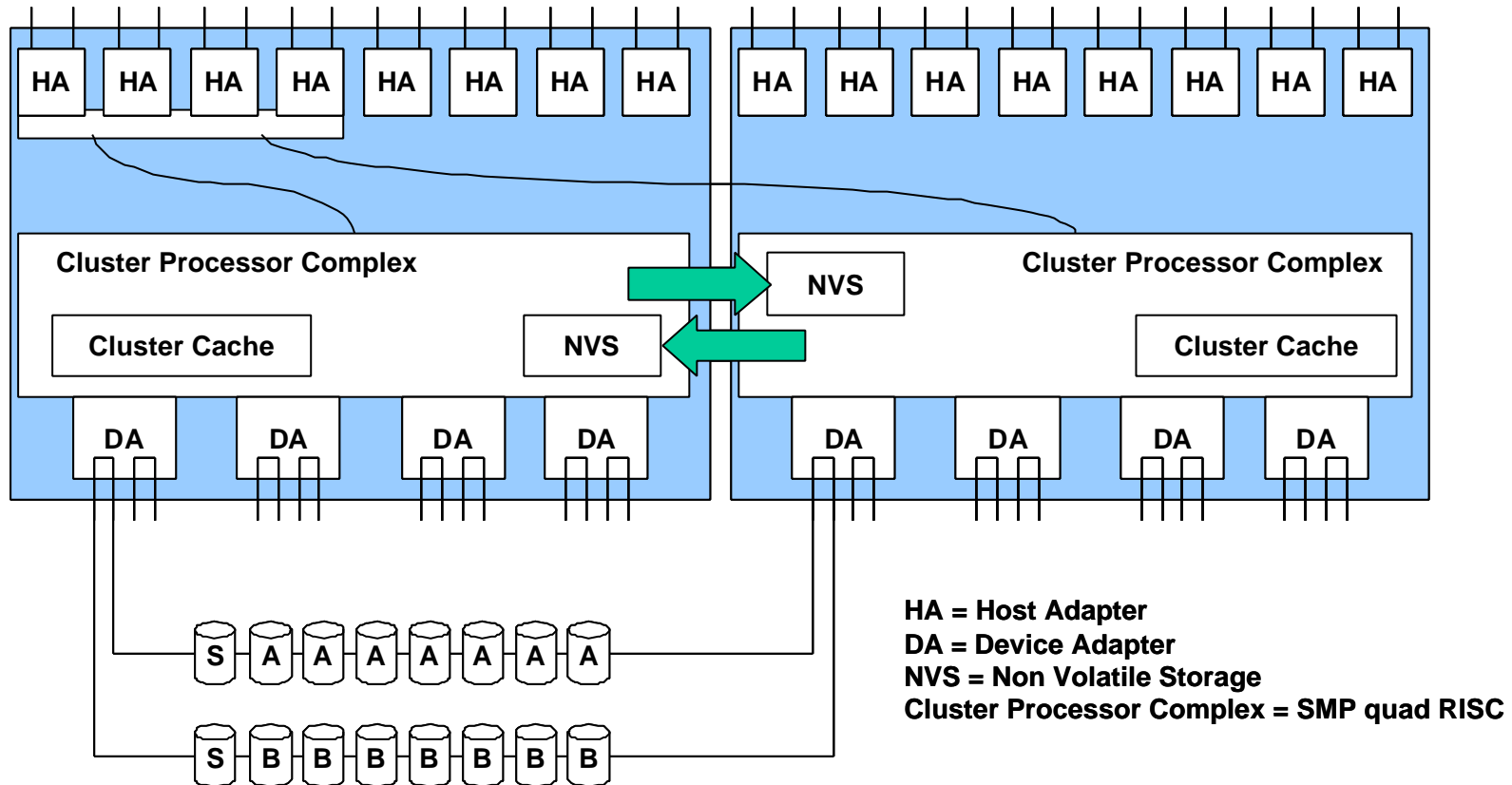
■ Implémentation NAS fondée sur les réseaux prédéfinis reprogrammables (Reprogrammable Gate Arrays).



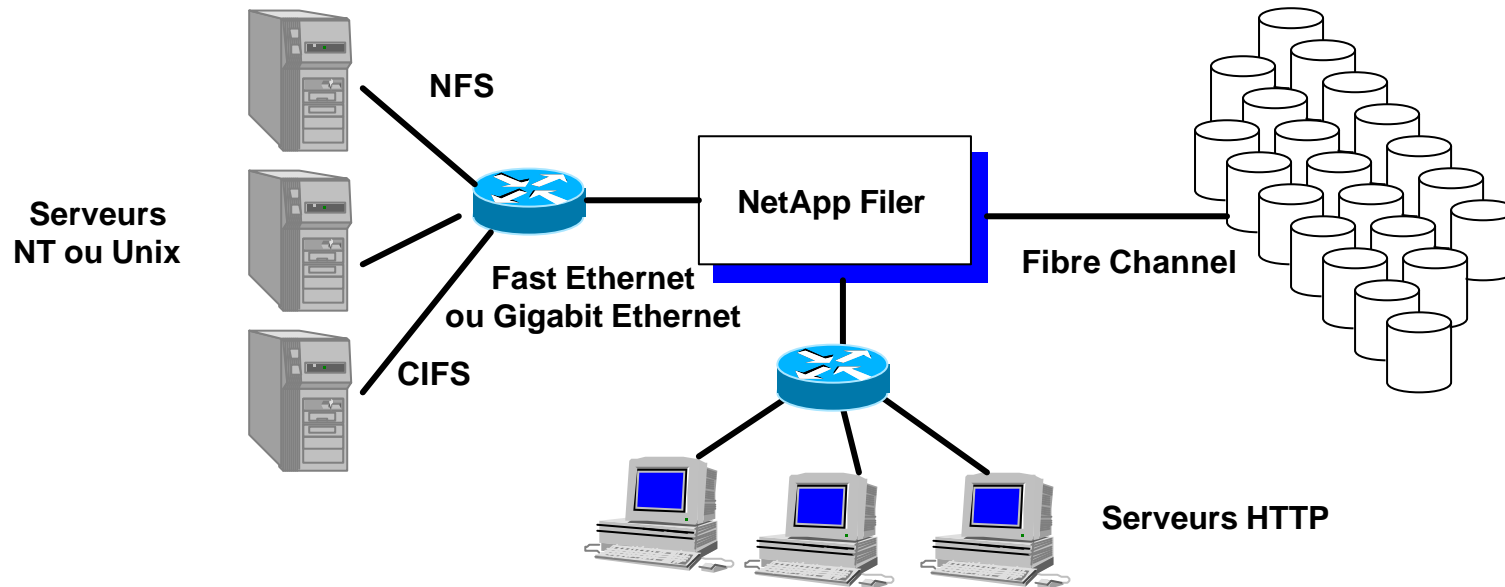
■ Celerra File Server/Symmetrix



■ Architecture ESS (Enterprise Storage Server)

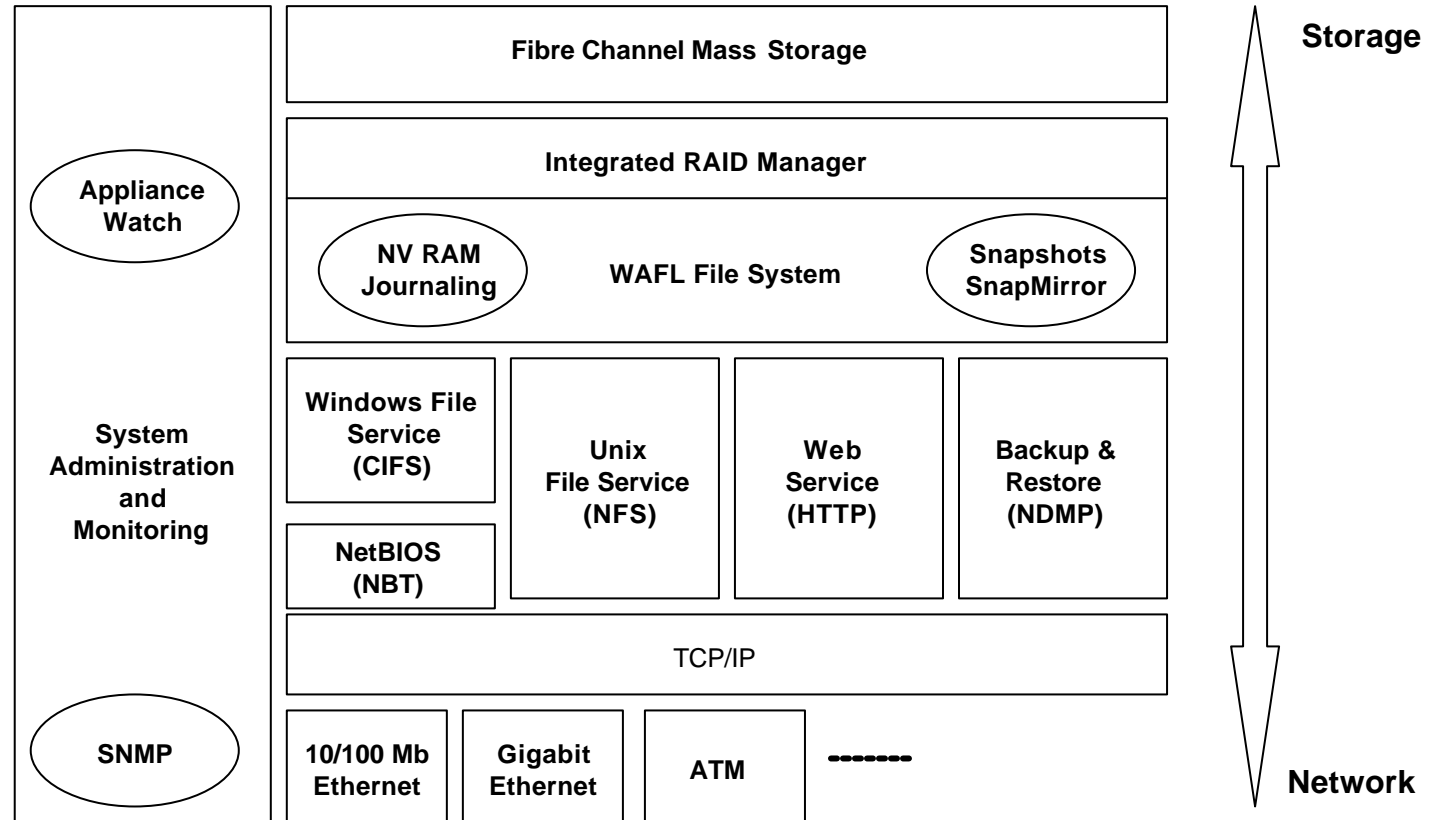


■ NetApp Filer



- Architecture matérielle classique fondée sur un processeur Alpha

■ Architecture du logiciel



WAFL = Write Anywhere File Layout (optimisé pour l'écriture)
CIFS = Common Internet File System

Note : Le WAFL est un « Log Structured File System » qui supporte la sauvegarde instantanée (snapshot)

■ Création de Snapshots (rappel)

