Base du stockage réseau sur IP

Besoin :

- croissance → facteur de 2 tous les 18 mois (Loi de moore)
- Interopérabilité & évolutivité : étendre, copier, déplacer un volume de stockage
- Continuité de service
- Performances

Objectifs :

- Manageability (copier et déplacer sans que l'utilisateur ne se rende compte)
- scalability (capacité de mise à l'échelle)
- availability (disponibilité)

Différents modes d'accès au stockage :

- Bloc (DAS, SAN) (on ne sait pas ce qu'on manipule, trame ICSI)
- Fichier (NAS) (Les systèmes de fichier réseau)
- Objet (métadonné plus importante que les données, profil de compte netflix …)

Type de stockage :

DAS : Direct Attached Storage (BLOC) Stockage machine

Usages : hôtes/Vms, transactions/SGBD

ISCSI -> trames SCSI + TCP

SAS : Serial Attached Storage | SCSI (format de la trame dans le SAS ou fibre channel FC) (un seul file)

Avantages : performances

Inconvénients : Dépendance redondance et haute disponibilité matérielles (RAID) et nombre de disques par baie

NVMe : Non Volatile Memory Express (accès parallèle plusieurs files fil d'exécution)

Avantages : latence réduite, performances, facteur de forme M.2

Inconvénients : Coût/capacité, évolutivité des architectures réseau

Gros bénéfice en cas de vm ou conteneur (accès multiple au disque)

SAN : Storage Area Network (BLOC) réseau dédié, réseau spécifique DATA Center

FCOE = Fiber channel over ethernet

fibre channel (FC) -> trames

ISCSI -> trames SCSI + TCP

Avantages : Performance + évolutivité, Redondance des liens -> multipath

Inconvénients : Interface (HBA) & commutateurs spécifiques (FcoE), cout

San : 1 application utilise 1 volume de stockage

NAS : Network Attached Storage (FICHIER) le réseau vient en premier fichier au dessus de la pile de protocole

NFS pour UNIX et SMB pour MICROSOFT

Avantages : nombre de clients, coût / unité de stockage, tiering (classement des données, plus les données sont nécessaire plus elles sont mise en avant (accés rapide) si nécessaire NVME sinon mois necessaire disque tournant),
caching (mise en cache de donnée),
déduplication (On stocke qu'une fois le fichier, calcul de hash au lieu d'avoir 50 fois l'objet présent sur le stockage il le sera qu'une fois utilisé principalement pour l'archivage. en cas de VM si les mêmes fichiers sont partout un seul est garder car il est nécessaire partout),
multi tenancy (un même logiciel accessible à plusieurs client),
réplication (on réplique lol, on peut stocker à n endroit différent)

Inconvénients : Architecture réseau (QoS) → coûts
(+Grosse connection nécessaire d'out la Qos)

NAS : 1 dispositif de stockage partager par n application

Object Storage : Mode objet

Fichier + Métadonnées = Objet

System Metadata : nom, taille, permissions, horodatage

Custom Metadata: Genre, auteur, localisation, contexte sécurité

Stockage non structuré : mobiles + cloud

Lieu de donnée différent entre Métadonnées et le fichier

	DAS	SAN	NAS	Object Storage
Accès	Mode bloc	Mode bloc	Mode fichier	Mode Objet
Connexion	SAS ou PCIe	Ethernet	IP	IP
Performances	Très bonnes	Très bonnes	Bonnes	Moins bonnes
Limite des performances	Sous-système noyau	Commutation réseau	Système de fichiers	Protocoles (HTTP)
Augmentation de capacité	Arrêt du système obligatoire	Facile	Très facile	Très facile
Évolutivité/ Continuité de service	Faible	Moyenne	Élevée	Très élevée (*cloud*)

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

Un profil d'accès pour chaque application :

- Ratio lecture / écriture
- Ratio données / méta-données
- Accès aléatoires ou séquentiels
- Efficacité de la compression
- Taille unitaire de bloc
- Commandes d'accès asynchrones ou par lots

Nature des accès au stockage
The I/O Blender effect

Virtualisation
    Profils hétérogènes -> VMs/ Conteneurs
    Un même réseau de stockage

Flux d'entrées/sorties aléatoires
    Flux d'une application -> séquentiels
    Flux des instances VMs/Conteneurs -> aléatoires

Accès aléatoires ou séquentiels
Relation entre Hyperviseurs et réseaux de stockage contraintes
    Intégrité -> Raid
    Réplication -> verrous LUNs

Protocole de Stockage :
ISCSI : protocole de stockage mode bloc qui permet d'accéder aux disques sur un réseau TCP/IP

NFS : protocole de stockage mode fichiers (répertoires, fichiers)

SMB : protocole de stockage mode fichiers (répertoires, fichiers). SMB direct est une variante d'accès direct entre application et réseau.

Objet :
CEPH : est un protocole ouvert conçu pour fonctionner sur du matériel de base. Basé sur l'algo CRUSH (Controlled Réplication Under Scalable Hashing) qui s'assure que les données sont réparties uniformément.

Swift : est le modèle OpenStack de stockage objet redondant et évolutif. Il peut utiliser un dispositif de stockage unique et est compatible avec Amazon S3.

Amazon S3 : est une solution de stockage en ligne Amazon Web Services. Il propose un accès au stockage via des web services tels que REST, SOAP et BitTorrent.

ISCSI en détails :

- Mode bloc au dessus de TCP
- Multi Chemins et redondance (MPIO)
- Réseau dédié au stockage (VLAN)
- Mode sans perte :
  - contrainte de latence (oversubscription)
  - solution logicielle (coût CPU plus élevé)
  - commutateur avec fonctions Data Center Bridging (DCB)

Volumes logiques : Logical Volume Manager (LVM)

Gestionnaire de périphérique mode bloc au niveau système
- Partition ou tout type de périphérique de stockage

Vus système homogène
- N Périphériques physiques vus comme un périphérique logique

Analogie entre volume et partition
- Formatage et création d'un système de fichiers

Changements dynamiques de configuration
- Snapshots, redimensionnement, extension, déplacements

Exemple pratique :

Combinaison entre DAS et SAN avec RAID1

- Hôte avec le rôle ISCSI target :
  - DAS → Fichier ou stockage local
  - SAN → Périphérique cible ISCSI
- Hôte avec le rôle ISCSI initiator :
  - SAN → nouveau volume de stockage réseau
  - DAS → volume réseau vu comme un stockage local
- Tolérance aux pannes sur l'hôte initiator :
  - RAID1 : réplication synchrone entre stockage local et stockage réseau.
  - LVM : snapshot entre disque système et tableau RAID1

Système de fichiers réseau & stockage objet

Un système de fichier réseau fournit une abstraction au système d'exploitation tandis que le stockage objet fournit une abstraction à une application. `rien que cette phrase fait déjà bugger`

VFS → Virtual File System
FUSE → Système de fichiers dans l'espace utilisateur
RPC → appels de procédures distants
NFS → Network File System
SMB → Server Message Bloc
CEPH → Stockage objet

Avantages :

Partage :

Fichier disponibles à grande échelle → nombre important de clients avec systèmes divers.
Partage depuis un point unique : diminution du nombre de copies et moins d'incohérences.

Gestion des disques :

système NAS → pas de file d'attente unique par périphérique en mode bloc.
Extension facile des volumes par ajout au serveur NFS → LVM

Architecture :

Gestion des verrous et des files d'attentes côté serveur → côté client avec les solutions SAN
Snapshots → au niveau système de fichier (LVM)
Sauvegardes → au niveau serveur NFS uniquement

Inconvénients :

Limitation technique

Lancement systeme impossible sur NAS
Coût CPU plus important côté hyperviseur -> logiciel de communication avec le serveur NAS

Performances

Latence possible -> usages applications transactionnelles
Pas de multipath avec un système NAS

Architecture

Manque de nouvelles fonctions liées à la virtualisation
Contraintes de sécurité -> autorisations d'accès depuis les clients
Accès sensibles aux ruptures de connexions réseau

Réseau IP frontal	Réseau de stockage dorsal
hote <-> hote application <-> système de fichiers client <-> serveur NFS / SMB NAS	hote <-> stockage système de fichiers <-> périphérique application <-> périphérique VFS noyau SAN

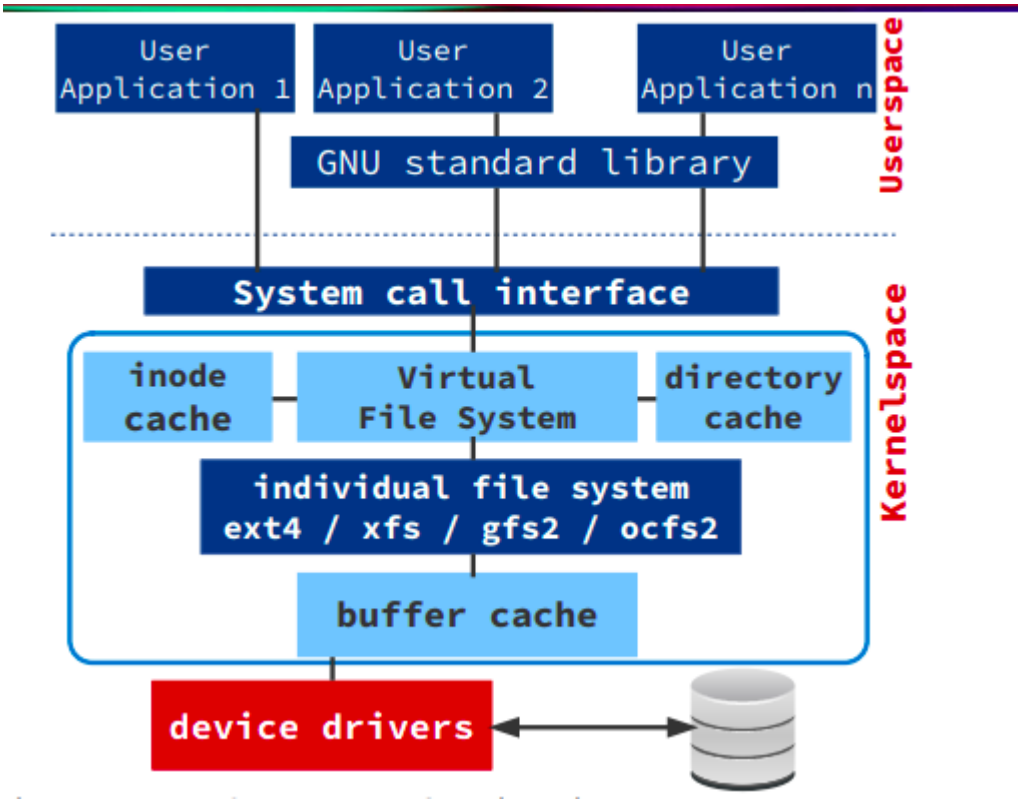
Système de fichiers virtuel (VFS)

Qu' est-ce qu'un système de fichiers virtuel (VFS)

- Un service du noyau
- Une interface entre les applications ou processus et les dispositifs de stockage
- Une bibliothèque standard

Quel est le but du VFS

- Un accès transparent au stockage vis-à-vis des utilisateurs et/ou des applications
- Une gestion uniforme du contrôle d'accès aux fichiers et répertoires
- Un système de nommage cohérent entre fichiers locaux et réseau
- Des performances d'accès uniformes -> cache



Système de fichiers dans l'espace utilisateur

Étape intermédiaire avant le stockage objet -> bibliothéqie FUSE

Portabilités entre noyaux

Interface de programmation
- Tous les langages
- Toutes les applications

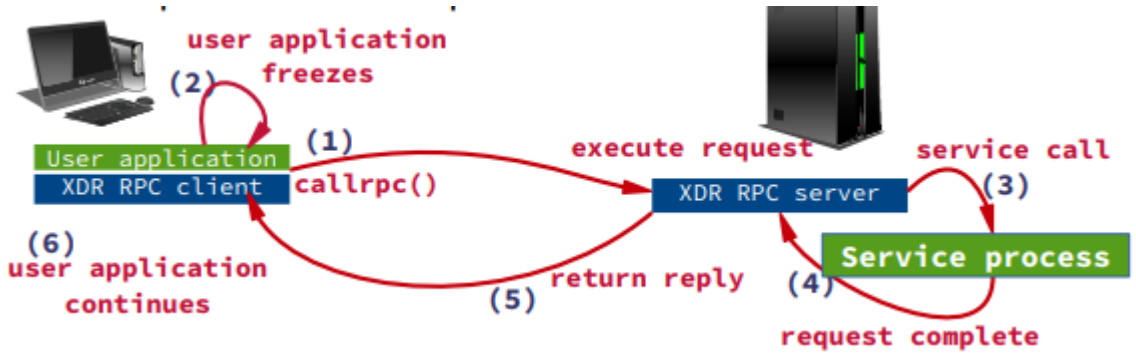
Avantages
- Performances
- Mise au point dans l'espace utilisateur

Inconvénient
- Métadonnées système uniquement

Appel de procédure distante :

Remote Procedure Call (RPC) :
- extension des appels de procédures locales
- exécution de code sur un hôte distant

Séquencement :



NFSv2 (1983) :
- Protocole de transport UDP uniquement réseaux locaux
- Performances médiocres en écriture

NFSv3 (1995) :
- Exportation de systèmes de fichiers POSIX 64bits
- Protocoles de transport UDP et TCP toujours avec multiplexage de ports
- Meilleures performances en écriture

NFSv4 (2003 puis maj en 2015) :
- Réduction des temps de latence
- Communication sur un port unique : TCP/2049 (filtrage plus simple)
- Appels de procédures groupées → compound RPC
- Chiffrement des flux (kerberos et SKPM LIPKEY)

Fonctions principales :

- Système de fichiers distribué → accès aux fichiers distant identiques aux accès locaux
- Modèle client / serveur :
  - serveur → répertoires locaux accessibles aux clients
  - clients → montage des répertoires distants
- Hiérarchique par nature → chaînes de répertoires et fichiers

Usages :

- Partages de répertoires : utilisateurs, données et applications exécutés localement
- Bases de données et VM datastores
- VMware supporte NFSv4.1 en mode client
- Amazon supporte NFSv4.0 à travers AWS Elastic File System (EFS)

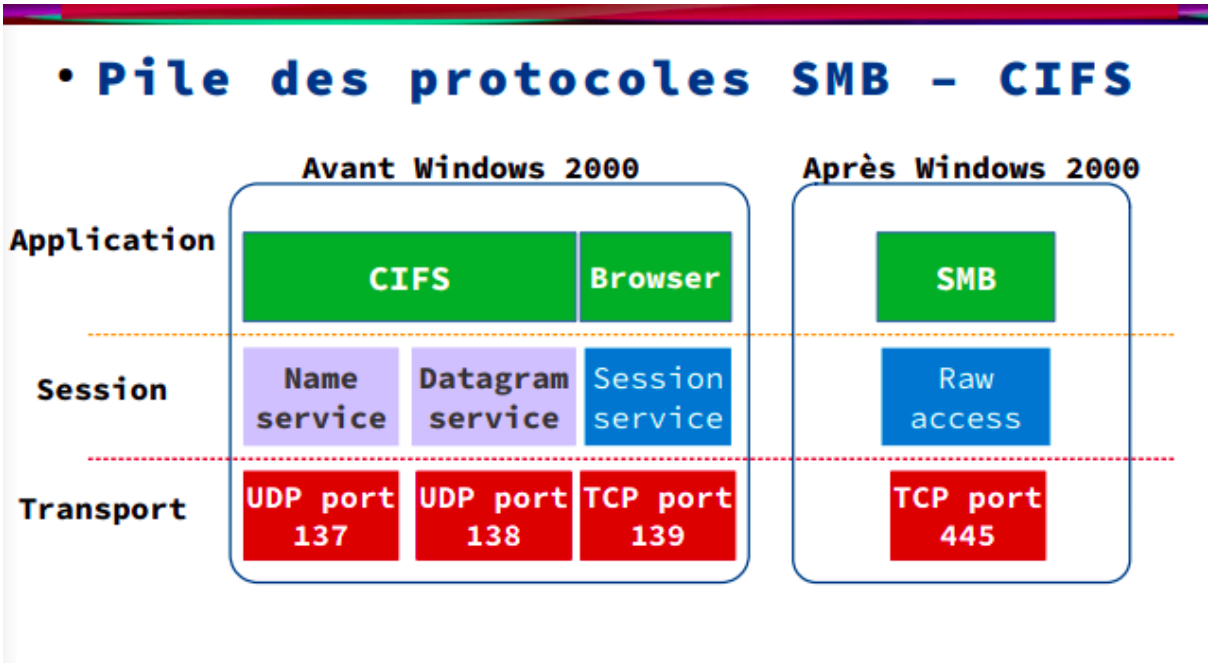
Évolutions du protocole :

- Paralle NFS (pNFS)
  - Agrégation de serveurs NFS autonomes
  - Relation entre client et serveur de type point-à-point
  - Client → lecture ou écriture d'un fichier sur tous les serveurs
  - Serveur → autorisation d'accès au fichier
  - Pagination (stripe map) du fichier fournie au client
  - Plus de goulot d'étranglement car plus de relation p2p avec le serveur
  - Gestion améliorée : répartition de charge entre clients et serveurs, espace de nommage unique préservé.

Server Message Block (SMB)
Fonctions principales
Mode de partage standard des systèmes Microsoft
SMB 3.0 -> windows server 2012 6> optimisé pour application serveur
Améliorations SMB 3.0
- Performances -> clusters
- Tolérance aux pannes
- Sécurité -> algorithmes AES

Usages
Intégration Hyper-V et produits Windows Server
Bases de données SQL Server

SMB Direct
Remote Direct Memory (RDMA)
Canaux d'accès “directs” de la mémoire d’un hôte à l’autre à travers le réseau -> Hyper-V
SMB Multichannel
Distribution des flux entre interfaces multiples

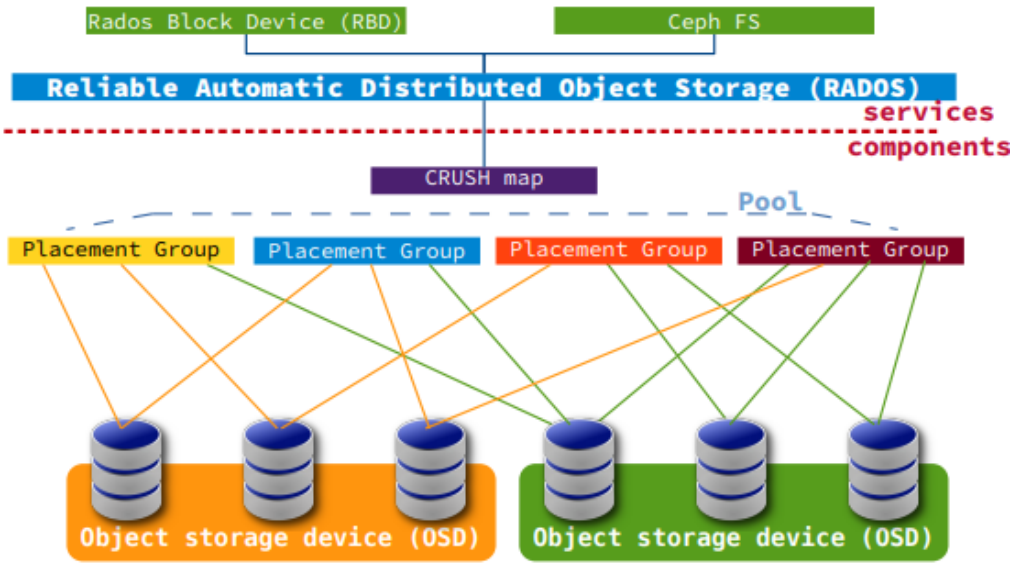


Object Storage
Stockage en mode objet
Trouver/chercher des données à partir d'expressions rationnelles
Cloud storage -> transformation en base de données
Plus de cloud se développe -> plus on peut trouver d'informations
Plus les métadonnées sont importantes -> plus on peut formuler de requêtes complexes (Big Data)
Exemple -> recherche de données similaires à l'aide de la classification des métadonnées

Stocker et récupérer un objet
Un serveur stocke un objet sur un seul noeud -> algorithmes de sélection basés sur la topologie
Politique de stockage -> algorithmes basés sur l'occupation des noeuds, les performances et les métadonnées système
Schémas de nommage cohérent -> accès global
Réplication -> entre noeuds et/ou sites en fonction des lectures

---

Architecture CEPH :



Object Storage Devices (OSDs) :
- Analogie avec les volumes de stockage
- OSDs multiples → RAID inutile

Pools :
- Regroupement de placement groups
- Stratégie de performance → tiering

CRUSH maps :
- Distribution des objets vers les OSDs
- Correspondance par pool

RADOS :
- Transformation → données / objets → stockage partagé performant

RDB :
- Périphérique de type bloc → découpage en stripes

Synthèse :

NAS :

- Fichiers → vue du système d'exploitation
- Partage réseau avec un bon rapport performance / coût

Stockage objet :

- objets → vue des applications
- Métadonnées → recherches et requêtes complexes
- Algorithmes / stratégies de distribution du stockage à grande échelle → cloud