

Gestió d'Infraestructures per al Processament de Dades

Emmagatzemament

Remo Suppi.

Departament Arquitectura d'Ordinadors i Sistemes Operatius

UAB (Remo.Suppi@uab.cat)

**Què
veurem?**

Emmagatzemament

Arquitectura

RAID

DFS

L'emmagatzematge

L'emmagatzematge és una de les infraestructures crítiques pel seu factor de coll d'ampolla com per ser la primera causa (55%) de fallades (server downtime ventilació 8%, memòria 5%, alimentació 28% i processador 4%).

Tecnologia xarxes d'emmagatzematge:

DAS (Direct Attached Storage): El medi està directament connectat a l'ordinador mitjançant un cable. Les peticions són per blocs o sectors.

NAS (Network Attached Storage): El dispositiu resideix en una xarxa que pot ser compartida amb un altre trànsit. No utilitza accés block I / O, sinó file I / O (peticions de més alt nivell -arxiu, desplaçament i nombre de bytes-, ja no es demanen sectors d'un disc directament). És el NAS (el seu SO) el que tradueix file I / O a block I / O.

NAS Gateways: igual funcionalitat de l'NAS però sense emmagatzematge en disc integrat. El GW fa de traductor de file I / O rebudes amb protocols com NFS a SCSI block I / O.

SAN (Storage Area Networks): L'emmagatzematge resideix en una xarxa dedicada. Les peticions d'E / S fan referència a blocs o sectors d'un dispositiu determinat. El concepte de SAN és independent de la xarxa que hi hagi per sota.

L'emmagatzematge: Tecnologia

Tots els dispositius informàtics tenen algun tipus de memòria d'emmagatzematge que utilitzen per emmagatzemar informació de forma estàtica. Hi ha tres tipus de tecnologies: magnètic, òptic, estat sòlid

Magnètic: codifica les dades en patrons de polaritat magnètica positiva i negativa en algun medi magnètic com el metall orgànic. Dispositius mecànics amb plats que són de material òxid de metall i giren a gran velocitat. Habitual GB/TB i diferents interfaces: **IDE**, **ATA**, **SCSI**, SATA (600 MB/s), SAS.

Òptic: utilitza làser per llegir o escriure dades. Es requereixen dues coses per a la tecnologia d'emmagatzematge òptic, una és el suport òptic (discs) i unitats òptiques. Tipus: CD (Compact Disc, 700MB), DVD (disc versàtil digital, 4,7 o 8,5 DVD-DL), BD (Blue-Ray Disc, 25, 50, 100, 128 GB, làser blau). Poden ser de lectura o de lectura escriptura (baixa velocitat)

Estat sòlid: fa servir memòria no volàtil que pot retenir les dades quan s'apaga l'alimentació. Aquesta tecnologia no té parts mòbils. L'emmagatzematge en estat sòlid avui en dia es és més car, però és més lleuger, ràpid, silencios i eficient que els anteriors. Tipus: Flash USB, targetes de memòria Flash (CompactFlash, Secure Digital (SD), MemoryStick i microSD), unitats d'estat sòlid (SSD).

SSD interfícies: **SAS-3 (12 Gbit/s)**, SATA 3.0 (6.0 Gbit/s), **PCI Express 3 (31.5 Gbit/s)** M.2 (6.0 Gbit/s for SATA 3.0 31.5 Gbit/s for PCIe 3), U.2 (PCIe 3), Fibre Channel (128 Gbit/s), **USB (10 Gbit/s)**

NVM Express (NVMe) especificació oberta per non-volatile storage media attached via PCI Express (PCIe) bus.



NAS: Network Attached Storage

SAN: Storage Area Network

L'emmagatzematge

Els **NAS** son generalment repositoris que fàcilment es connecten a la xarxa corporativa sense necessitat d'utilitzar un servidor.

La tecnologia NAS permet accés compartit a **fitxers**, des d'un servidor especialitzat amb HW i SW optimitzat, ús de protocols estàndard (NFS o Common Internet File System) i la seva instal·lació és molt fàcil.

Inconvenients: ús compartit de la xarxa de producció que conté trànsit de transaccions d'emmagatzematge des de les connexions SCSI paral·leles i tràfic dels usuaris.

És per això que si es fan backups sobre el NAS consumeix ample de banda dels servidors.

El **SAN** és un entorn d'emmagatzematge amb a **xarxa especialitzada (generalment fibra)** que permet un accés ràpid i fiable entre servidors i recursos d'emmagatzematge independents o externs **a nivell de bloc** de manera que el dispositiu d'emmagatzematge apareix localment connectat a sistema operatiu.

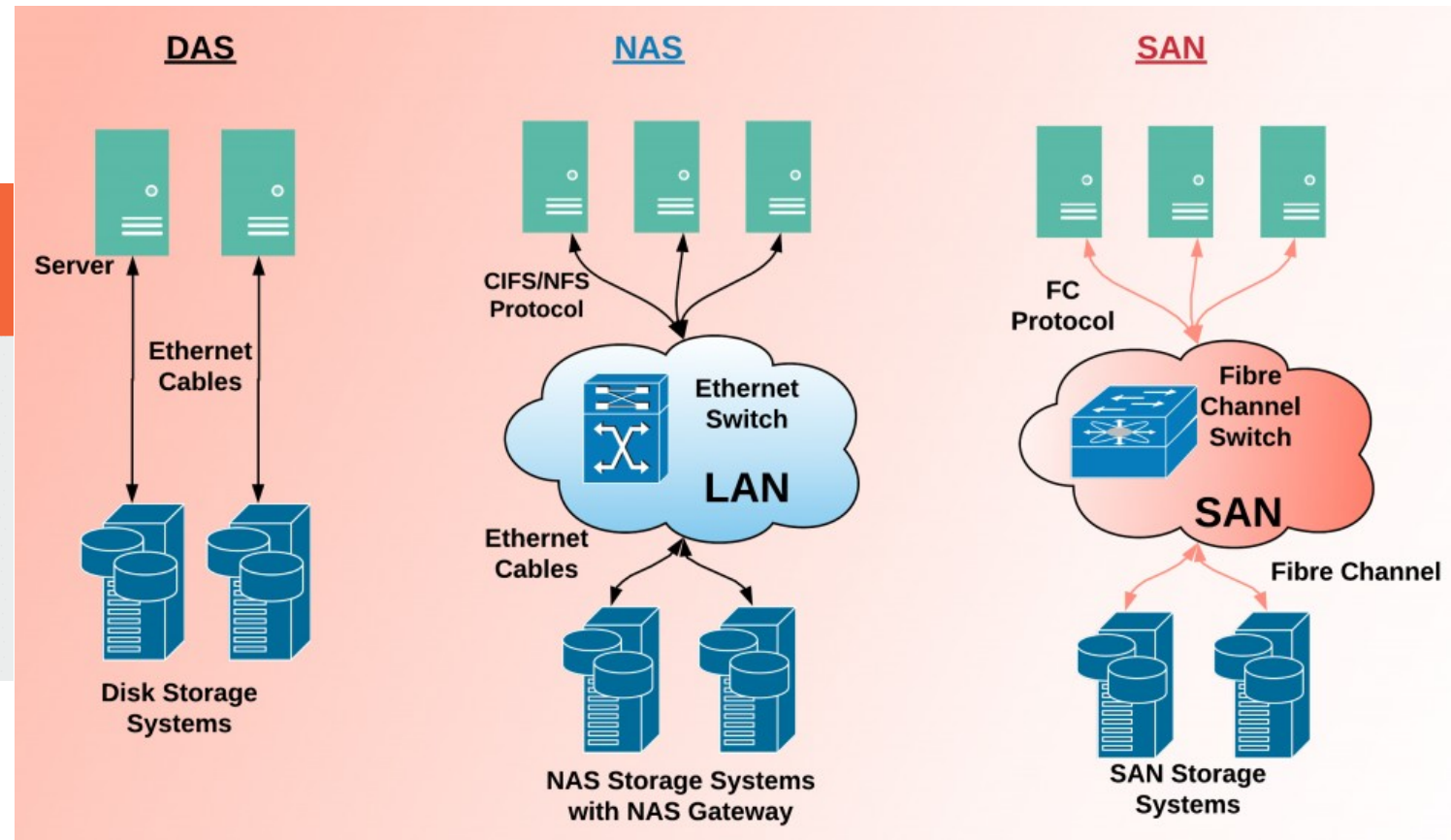
Una SAN **no proveeix *file abstraction* sinó operacions a nivell de blocs només** (coneguts *SAN filesystems* o *shared disk file systems*).

Moltes xarxes d'emmagatzematge usen el protocol SCSI per a la comunicació entre servidors i discos però es poden utilitzar un *mapping layer* a altres protocols per adequar-los a la xarxa.

L'emmagatzematge

NAS vs SAN

NAS (Network attached storage)	SAN (Storage area network)
<ul style="list-style-type: none">• File level data• Primary Media: ethernet• I/O Protocol: NFS/CIFS• NAS appears to OS as a shared folder• Inexpensive• Dependent on the LAN• Requires no architectural changes	<ul style="list-style-type: none">• Block level data• Primary Media: fiber channel• I/O Protocol: SCSI• SAN appears to OS as attached storage• Expensive• Independent of the LAN• Requires architectural changes



L'emmagatzematge

SCSI: Small Computer System Interface

És un conjunt d'estàndards per connectar i transferir físicament dades entre ordinadors i dispositius perifèrics.

Els estàndards SCSI defineixen ordres, protocols, interfícies elèctriques, òptiques i lògiques. SCSI s'utilitza més habitualment per a unitats de disc dur i unitats de cinta, però pot connectar una àmplia gamma d'altres dispositius, inclosos escàners i unitats de CD, tot i que no tots els controladors poden gestionar tots els dispositius.

L'estàndard SCSI defineix conjunts d'ordres per a tipus de dispositius perifèrics específics i en teoria, es pot utilitzar com a interfície per a gairebé qualsevol dispositiu ja que l'estàndard és altament pragmàtic.

Tipus de connexió:

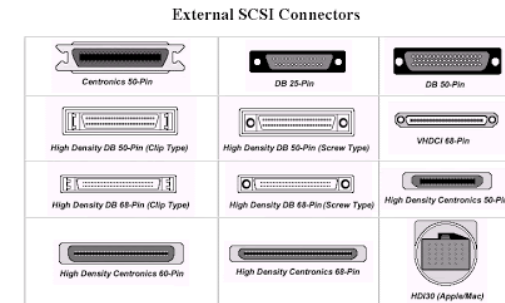
Interfície paral·lela SCSI: interns solen ser cintes, amb dos o més connectors de 50–, 68– o 80-pins i els externs apantallats (però no ho poden ser), amb connectors de 50 o 68 pins a cada extrem, depenent de l'amplada de bus SCSI específica admesa.

Canal de fibra: es pot utilitzar per transportar unitats d'informació SCSI, tal com es defineix en el protocol de canal de fibra per a SCSI (FCP). Aquestes connexions es poden connectar en calent.

Serial attached SCSI (SAS): SCSI connectat en sèrie (SAS) utilitza un cable d'alimentació i dades Serial ATA (SATA) modificat.

iSCSI (Internet Small Computer System Interface): utilitza connectors i cables Ethernet com a transport físic, però pot funcionar per qualsevol transport físic capaç de transportar IP.

USB Attached SCSI: permet als dispositius SCSI utilitzar el USB.



L'emmagatzematge

SCSI: Small Computer System Interface

L'actualitat de SCSI (V3)

Serial Attached SCSI (SAS), SCSI-over-Fiber Channel Protocol (FCP) i USB Attached SCSI (UAS)

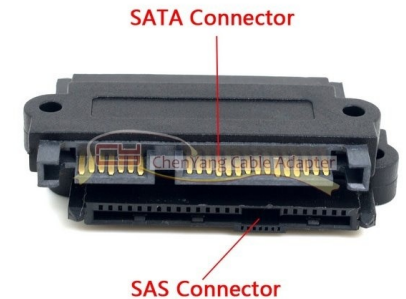
Han abandonat el tradicional SCSI paral·lel i utilitzen interfícies en sèrie que tenen moltes avantatges: velocitats de dades més altes, cablejat simplificat, abast més llarg, aïllament de fallades millorat i capacitat full-duplex (la principal raó no obstant és que les paral·leles a alta velocitat tenen gran quantitat de problemes causats pel cablejat i els connectors).

SCSI és popular en estacions de treball, servidors i aparells d'emmagatzematge d'alt rendiment. Gairebé tots els subsistemes **RAID** dels servidors han utilitzat algun tipus de disc dur SCSI (inicialment SCSI paral·lel, Fibre Channel provisional, recentment SAS), tot i que diversos fabricants ofereixen subsistemes RAID basats en SATA com una opció més barata.

A més, SAS ofereix compatibilitat amb dispositius SATA, creant una gamma d'opcions molt més àmplia per als subsistemes RAID. En lloc de SCSI, els ordinadors i ordinadors portàtils moderns solen utilitzar interfícies SATA per a les unitats de disc dur interns, amb NVMe sobre PCIe guanyant popularitat, ja que SATA pot ser un coll d'ampolla les modernes unitats d'estat sòlid.

USB Attached SCSI (UAS) o USB Attached SCSI Protocol (UASP): protocol que s'utilitza per moure dades a i des de dispositius d'emmagatzematge USB, com ara discs durs (discs durs) i unitats d'estat sòlid (SSD) UAS depèn del protocol USB i utilitza el conjunt d'ordres SCSI estàndard i proporciona transferències més ràpides en comparació amb els controladors de transport massiu d'emmagatzematge massiu USB (Bulk-Only Transport, BOT) més antics.

UAS es va introduir com a part de l'estàndard USB 3.0, però també es pot utilitzar amb dispositius que compleixin l'estàndard USB 2.0 més lent, suposant l'ús de maquinari, firmware i controladors compatibles.



L'emmagatzematge

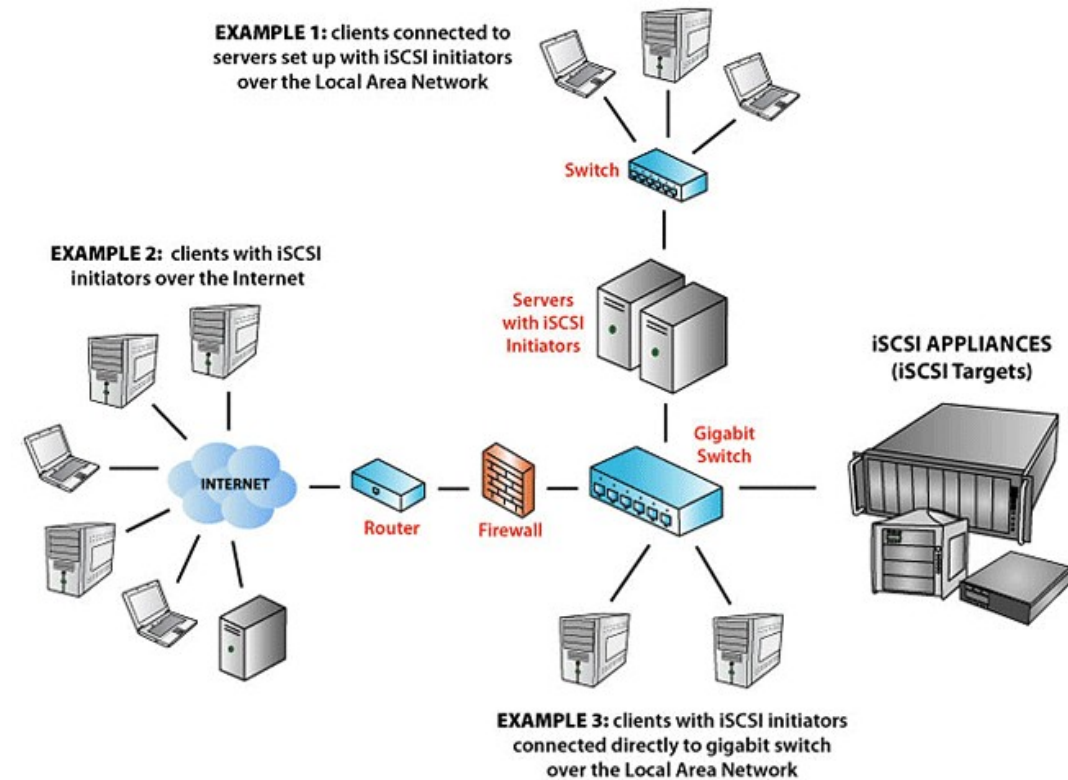
estàndard que permet l'ús del protocol SCSI sobre redes TCP / IP. iSCSI és un protocol de la capa de transport definit en les especificacions SCSI-3.

L'adopció del iSCSI en entorns de producció i clústers té gran acceptació per les avantatges de Gigabit Ethernet. El seu preu és molt competitiu i resulta una alternativa de solucions **SAN** basades en fibra.

El protocol iSCSI utilitza TCP/IP per a les seves transferències de dades i solament requereix una interfície Ethernet senzilla i senzilla per funcionar. Això permet una solució d'emmagatzematge centralitzat de baix cost sense la necessitat de realitzar inversions costoses ni patir les habituals incompatibilitats associades a les solucions canal de fibra per a redisseny d'àrea d'emmagatzematge.

Els crítics argumenten que aquest protocol té un pitjor rendiment que el canal de fibra ja que es veu afectat per la sobrecàrrega que generen les transmissions TCP/IP (capçaleres de paquets). No obstant això les proves que s'han realitzat mostren un excel·lent rendiment de les solucions iSCSI SANs, quan s'utilitzen enllaços Gigabit Ethernet.

iSCSI: Internet Small Computer System Interface

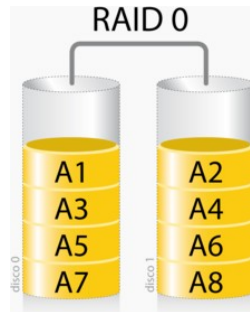


RAID (Redundant Array of Independent Disks)

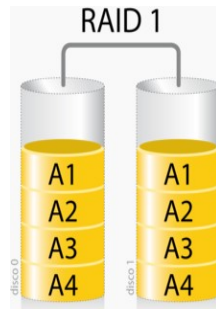
Millora les prestacions fent operacions de lectura/escriptura d'un conjunt de discos al mateix temps i la fiabilitat la incrementa incloent paritat en la informació o replicant la informació sobre múltiples disc de la matriu.

RAID pot ser de tipus maquinari amb un controlador específic o mitjançant programari al host.

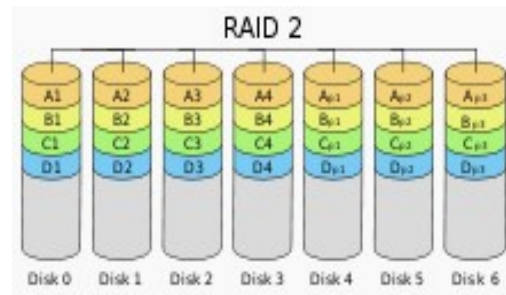
Tipus més comuns de Raids són: (estàndard) 0 (Data Striping), 1,2,3,4,5,6, 5e i 6e (nested) 0 + 1, 1 + 0, 30, 100, 50 (existeixen també nivells de Raid propietaris que tenen variacions sobre els estàndard).



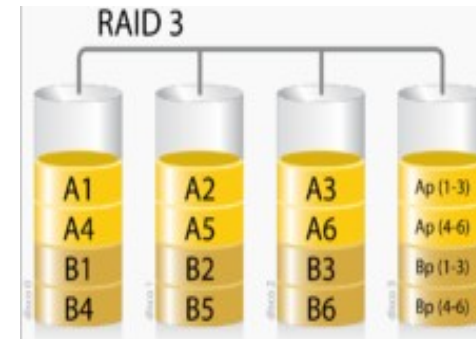
Divisió de dades, sense paritat. Cost baix. Size = < disco



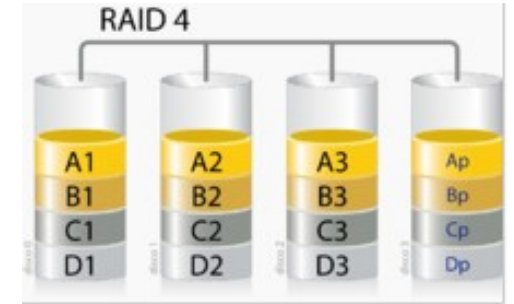
Mirall. Sense paritat. Costo baix. Size = < disco



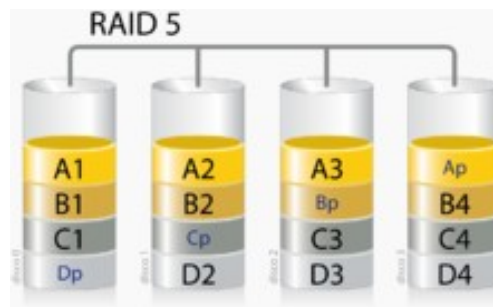
No es fa servir. División en bits. **in**Costo = 32 discos dades + 7 paridad.



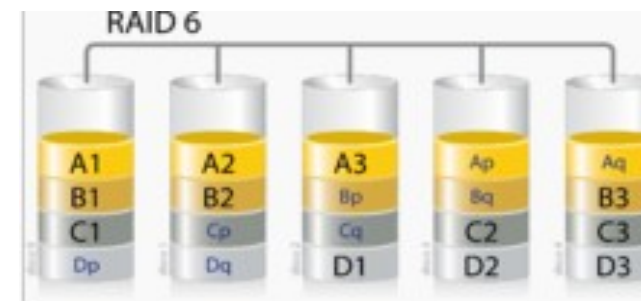
Poc utilitzat. Divisió en bytes. Paritat.. **in**Costo = 32 discos dades + 7 paritat. NO operacions en paral·lel



Divisió en blocs + disc de paritat. Mínim 3 discos. Lectura A1, B2 en paral·lel, escriptura seqüencial



Divisió en blocs i paritat distribuïda. Possibilitat de recuperar blocs disc amb errors. Mínim 3 discos. Lectura x 2



Ídem R5 però amb diferent algoritme de paritat (Reed-Solomon). Mínim 4 discos ja que fa servir doble paritat.

Crear un RAID software: **mdadm** és l'ordre Linux que es fa servir per a administrar dispositius RAID de programari. El nom deriva de md (múltiple device) i de administers (reemplaça a l'anterior mdctl).

Instal·lar: **apt-get install mdadm**

Preparar els discos: **fdisk** con una partició tipus 0xfd-, i després **mkfs.ext4 /dev/sdi1**

Per crear un raid 5 anomenat */dev/md0* con 3 dispositius i chunk size of 16384: (típicament un gran chunk és millor per gran arxius grans, default es 512)

mdadm --create --level=5 [--chunk=16384] --raid-devices=3 /dev/md0 /dev/sdb1 /dev/sdc1 /dev/sdd1

I crear el *filesystem* abans de muntar-ho.

Per muntar un array que no està a l'arxiu de configuració:

mdadm --assemble /dev/md0 /dev/sdb1 /dev/sdc1 /dev/sdd1

Per crear l'arxiu de configuració per a que arranqui al iniciar l'ordinador:

mdadm --detail --scan >> /etc/mdadm/mdadm.conf

(Crear el filesystem en */dev/md0* antes de montarlo)

RAID (Redundant Array of Independent Disks)

Per afegir un nou disc i fer créixer l'array (després s'haurà de fer créixer el filesystem amb **resize2fs /dev/md0**):

mdadm --add /dev/md0 /dev/sde1 (aquest disc quedarà com spare)

mdadm --grow /dev/md0 --raid-devices=4 (amb això veurem que passa d'spare a actiu)

Marcar un disc amb errors i treure'l de l'array (també és necessari quan volem canviar un disc)

mdadm --fail /dev/md0 /dev/sde1

mdadm --remove /dev/md0 /dev/sde1

Aturar l'array : (haurem de assegurar que primer el filesystem estigui desmuntat)

mdadm --stop /dev/md0

Iniciar l'array (després es podrà muntar):

mdadm --run /dev/md0

Obtenir info de l'array:

mdadm --detail /dev/md0

Monitoritzar l'arrays: (i obtenir mails de quan té errors)

mdadm --monitor --scan --mail=[email address] --delay=1800 &

És un sistema de fitxers d'emmagatzematge distribuït. Es fa servir en cloud, serveis streaming i xarxes de contingut (CDN). GlusterFS va ser desenvolupat originalment per Gluster, Inc. i després per Red Hat (adquisició en 2011, en 2012 es va integrar en Red Hat Storage Server en 2014 va comprar Inktank Storage (Ceph), i va canviar el nom de Red Hat Storage Server basat en GlusterFS a "Red Hat Gluster Storage".

Disseny: GlusterFS agrupa diversos servidors d'emmagatzematge a través d'interconnexió Ethernet o Infiniband en un gran sistema de fitxers de xarxa. Es tracta de programari lliure amb GPL/LGPL. GlusterFS es basa en un disseny *d'stackable user space* i funciona com client i servidor.

Normalment, els servidors proveïxen el 'maons' d'emmagatzematge, i el client, que es connecta a servidors amb un protocol personalitzat a través de TCP/IP, InfiniBand o Sockets Direct Protocol, crea volums virtuals compostos a partir de diversos servidors.

Per defecte, els fitxers s'emmagatzemen sencers, però també és possible separar-los en diversos volums remots. El client pot muntar el volum compost mitjançant un protocol natiu de GlusterFS mitjançant el mecanisme FUSE o per NFS v3.

GlusterFS proporciona fiabilitat i disponibilitat de dades a través de diversos tipus de rèplica: volums replicats i geo-replicació.

Els volums replicats garanteixen que hi hagi almenys una còpia de cada fitxer ens el 'maons', de manera que si un falla, les dades continuaran emmagatzemades i accessibles. La geo-replicació proporciona un model de replica *master/worker*, on es copien els volums en ubicacions geogràficament diferents. Això passa de manera asíncrona i és útil per a la disponibilitat en cas de que un centre de dades quedi fora de servei.

apt glusterfs-client glusterfs-server

gluster peer probe serverA descobreix el servidor

gluster peer status veure l'estat

gluster volume create myvol transport tcp serverA:/export/brick serverB:/export/brick crear el volum

gluster volume start myvol iniciar-lo

gluster volume info obtenir info

Si volem crear 2 repliques de cada arxiu:

gluster volume create myvol replica 2 serverA:/export/brick serverb:/export/brick force

Veure si s'ha creat bé: **gluster volume info**

Iniciar el volum: **gluster volume start myvol**

Crear un punt de muntatge i muntar el volum:

mkdir -p /mnt/gluster

mount -t glusterfs serverA:/myvol /mnt/gluster

dh -h

També es pot posar en el /etc/fstab:

serverA:/myvol /mnt/gluster glusterfs rw,noauto 0 0

Es provar amb:

for i in `seq -w 1 15`; do cp -rp /var/log/messages /mnt/gluster/test-\$i; done

Ordres útils:

gluster volume stop nom

gluster volumen delete nom

gluster peer detach serverA

gluster peer probe serverA

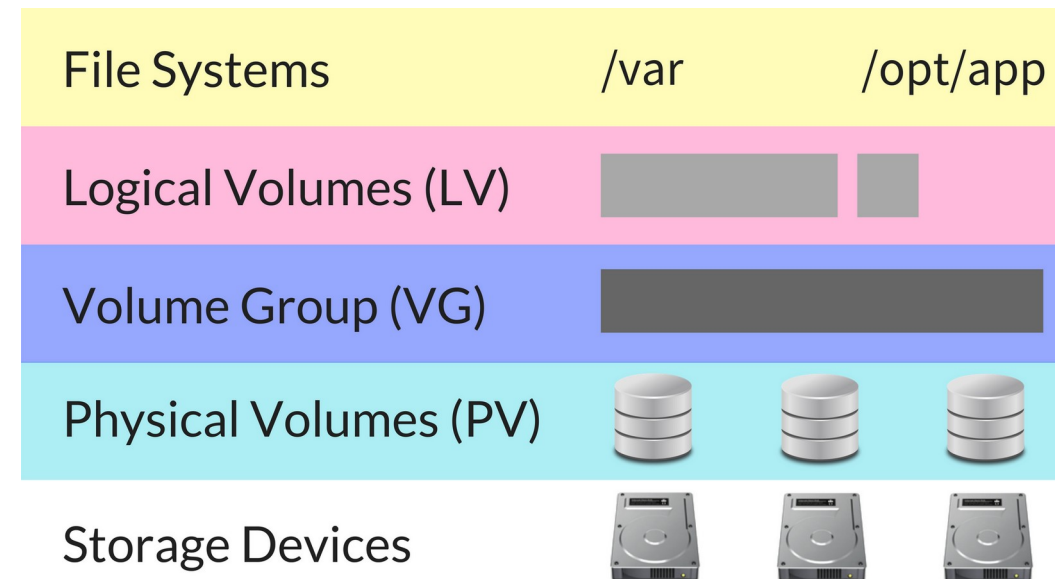
gluster peer status

LVM (Logical Volume Manager)

És un *framework* que proporciona una gestió de volum lògica per al nucli Linux separant la gestió física del disc de la lògica. El seu creador (H. Maues) va escriure el codi original el 1998 basat en les idees del gestor de volums de l'HP-UX.

LVM s'utilitza per als propòsits següents:

- Creació de volums lògics individuals de diversos volums físics o discs sencers (semblants a RAID 0), permetent canviar la mida del volum en forma dinàmica.
- Gestió de granges de discs durs permetent afegir i substituir discos sense temps d'inactivitat ni interrupció del servei, en combinació amb l'intercanvi en calent.
- En sistemes petits en lloc d'haver d'estimar la mida que pot necessitar una partició, LVM permet canviar la mida dels sistemes de fitxers segons sigui necessari.
- Realització de còpies de seguretat consistentes prenent instantànies dels volums lògics.
- Xifrar diverses particions físiques amb una sola contrasenya.



LVM (Logical Volume Manager)

Inicialització en 4 passos:

- Definir i inicialitzar els discos creant una partició en cada disc tipus **8e** (amb el fdisk)
- Definir i inicialitzar cada **Physical Volumes** (PV): **pvcreate /dev/device [/dev/device]**
- Definir el **Volume Groups** agrupant els PVs (VG): **vgcreate vg-name /dev/device [dev/device]**
- Inicialitzar els **Logical Volumes** sobre cada VG (LV): **lvcreate -L size -n lv-name vg-name**

Espot expandir (o reduir) un volum: a) **pvcreate /dev/nou_dev** b) **vgextend vg-name /dev/nou_dev**
c) (per reduir) **vgreduce vg-name /dev/dev**

```
sudo apt install lvm2
sudo apt install system-config-lvm
sudo pvcreate /dev/vdd1 /dev/vde1
sudo vgcreate myvol /dev/vdd1 /dev/vde1
sudo vgdisplay
sudo lvcreate -n mylogvol -L 10g myvol
sudo mkfs.ext4 /dev/myvol/mylogvol
mkdir /test
sudo mount /dev/myvol/mylogvol /test; df -h
sudo lvdisplay
sudo lvremove /dev/myvol/mylogvol
```

Resum:

lvchange	Change attributes of a Logical Volume.
lvconvert	Convert a Logical Volume from linear to mirror or snapshot.
lvcreate	Create a Logical Volume in an existing Volume Group.
lvdisplay	Display the attributes of a Logical Volume.
lvextend	Extend the size of a Logical Volume.
lvreduce	Reduce the size of a Logical Volume.
lvremove	Remove a Logical Volume.
lvrename	Rename a Logical Volume.
lvresize	Resize a Logical Volume.
lvs	Report information about Logical Volumes.
lvscan	Scan (all disks) for Logical Volumes.

Ceph és una plataforma software *open-source* de emmagatzemament que implementa *object storage* sobre clúster i proveix interfícies 3 in1 per: **object, block, file-level storage**.

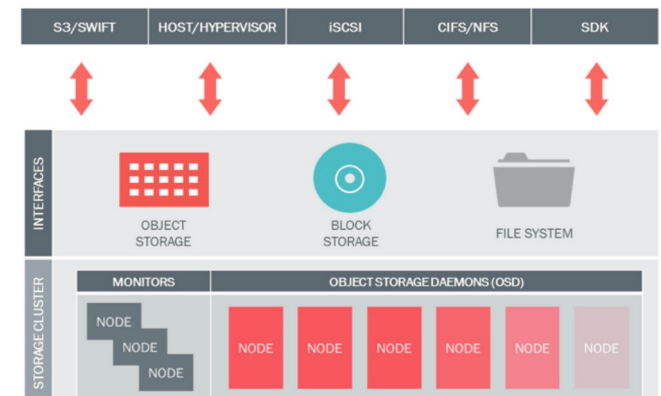
Ceph té com a objectiu principal un funcionament completament distribuït sense cap punt de fallada, escalable fins al nivell d'exabyte i de lliure accés.

Ceph replica les dades i les fa tolerants a fallades, utilitzant maquinari bàsic i que no requereix cap suport de maquinari específic. Com a resultat del seu disseny, el sistema és auto-reparable i auto-gestionable, amb l'objectiu de minimitzar el temps d'administració i altres costos.

Nivell de fitxer: habitual amb nom dels fitxers, pot portar metadades i organització carpetes de directoris i subdirectoris. Es fa servir en el NAS.

Nivell bloc: utilitzats a les SAN. Un bloc és un volum d'emmagatzematge en brut ple de fitxers que s'han dividit en trossos de dades d'igual mida. Un SO gestiona aquests volums i els pot utilitzar com a discs durs individuals. Habitual en bases de dades, servidors de correu electrònic, RAID, màquines virtuals.

Nivell objecte: emmagatzema dades en contenidors aïllats coneguts com a objectes. Es pot donar a un sol objecte un identificador únic i emmagatzemar-lo en un model de memòria plana. Facilita la cerca, és flexible en quan a les metadades i l'emmagatzematge remot. És altament escalable i transparent.



Suite de codi obert, que proporciona servei d'arxius i d'impressió a clients SMB/CIFS i pot compartir un sistema d'arxius de Linux amb Windows i viceversa i també impressores connectades a Linux o un sistema amb Windows.

Bàsicament Samba permet que un servidor Linux (o Unix) funcioni com un servidor d'arxius per a equips client que executen Windows.

SMB (Server Message Block) i CIFS (Common Internet File System) és un protocol que inclou SMB amb un sèrie de característiques addicionals.

apt-get install samba

L'arxiu de configuració: /etc/samba/smbd.conf

systemctl restart smbd

apt-get install smbclient

Creació de un recurs compartit entre W/MacOS y el servidor: Crear el directori i canviar el permisos

mkdir /home/share; chmod 777 /home/share

Editar l'arxiu de configuració /etc/samba/smb.conf

```
....
workgroup = WORKGROUP
....
interfaces = 127.0.0.0/8 10.0.0.0/24
bind interfaces only = yes
...
map to guest = Bad User
....
```

[Share]

```
path = /home/share
writable = yes
guest ok = yes
guest only = yes
create mode = 0777
directory mode = 0777
share modes = yes
```

Reiniciar el
servei:
systemctl restart
smbd

Provar localment:

smbstatus

smbclient -L localhost

smbclient '\\localhost\\Share' passwd

Provar externament: des de Windows (explorer) [\\ip_asignada_al_servidor](#)

Directorí restringit amb autenticació:

groupadd security

mkdir /home/security

chgrp security /home/security

chmod 770 /home/security

vi /etc/samba/smb.conf

[Security]

path = /home/security

writable = yes

create mode = 0770

directory mode = 0770

share modes = yes

guest ok = no

valid users = @security # solo el security group

Afegeixo usuari en Samba

smbpasswd -a jessie

usermod -G security jessie

systemctl restart smbd

Configure Samba Active Directory Domain Controller

Provar amb smbclient o amb W/MacOS i es podrà accedir amb usuari i passwd.

Cloud: AWS

Cloud: Azure

Cloud: Google

Object storage



Amazon Simple Storage Service (S3)

Object storage built to store and retrieve any amount of data from anywhere

File storage



Amazon Elastic File System

Scalable, elastic, cloud-native NFS file system



Amazon FSx for Windows File Server

Fully managed file storage built on Windows Server

Block storage



Amazon Elastic Block Store

Easy to use, high performance block storage at any scale



Amazon FSx for Lustre

Fully managed high-performance file system integrated with Amazon S3

Backup



AWS Backup

Centrally manage and automate backups across AWS services



AWS Snow Family

Physical edge computing and storage devices for rugged or disconnected environments

- **Azure Blobs:** A massively scalable object store for text and binary data. Also includes support for big data analytics through Data Lake Storage Gen2.
- **Azure Files:** Managed file shares for cloud or on-premises deployments.
- **Azure Queues:** A messaging store for reliable messaging between application components.
- **Azure Tables:** A NoSQL store for schemaless storage of structured data.
- **Azure Disks:** Block-level storage volumes for Azure VMs.

<https://docs.microsoft.com/en-us/azure/storage/common/storage-introduction>

OBJECT OR BLOB STORAGE

BLOCK STORAGE

ARCHIVAL STORAGE

FILE STORAGE

MOBILE APPLICATION

DATA TRANSFER

COLLABORATION

<https://aws.amazon.com/products/storage/>

<https://cloud.google.com/products/storage/>

UAB

Universitat Autònoma de Barcelona

Administració de sistemes GNU/Linux, 2016: <http://openaccess.uoc.edu/webapps/o2/handle/10609/60687>
Adm. Avançada, 2016: <http://openaccess.uoc.edu/webapps/o2/handle/10609/60685>

Glusterfs:

<https://docs.gluster.org/en/latest/Administrator-Guide/>

https://www.server-world.info/en/note?os=Debian_12&p=glusterfs&f=1