75-08 Sistemas Operativos
Lic. Ing. Osvaldo Clúa
2010

Facultad de Ingeniería
Universidad de Buenos Aires

Clustered File Systems

Clustered file system

- Es un file system que va ser accedido simultáneamente desde mas de un cliente.
  - En general NO es usado por los clusters.
  - Proveen un mecanismo de control de concurrencia y de serialización.
    - A nivel de bloques como RAID y SANs
    - A nivel de Archivo/Registro como en NASs.

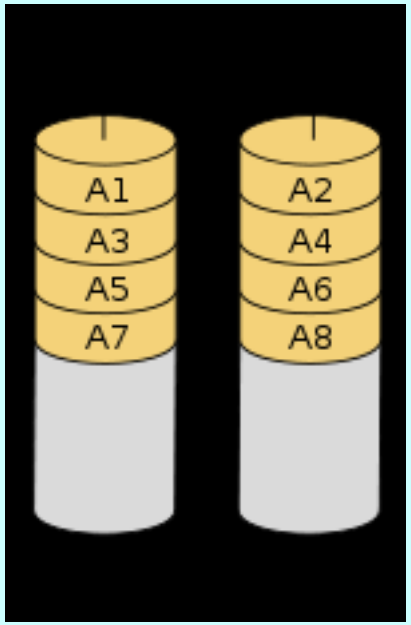
RAID

- Redundant Array of (Inexpensive/ Independent) Disks.
  - Concepto desarrollado por David A. Patterson, Garth A. Gibson, and Randy Katz en la University of California, Berkeley en 1987.
- Hoy es un término "paraguas" para replicar y dividir datos entre varios discos.
  - Pero vistos como un solo disco por el Sistema Operativo.

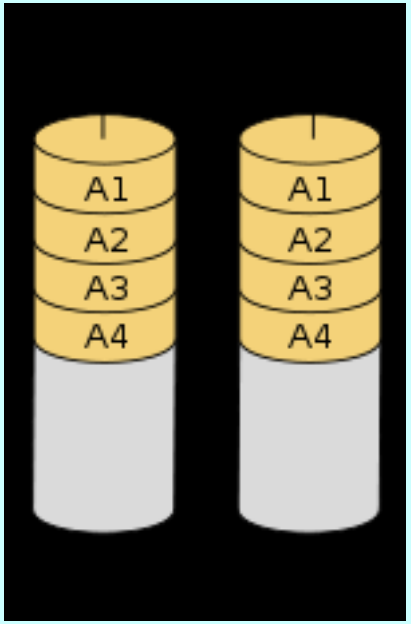
Principios del RAID

- Combinar varios discos físicos en una única unidad lógica.
  - Por Software o Hardware
- Provee varios esquemas de:
  - Mirror o redundancia de datos
  - Stripping o distribución de bloques de datos
  - Corrección de errores.

Niveles de RAID

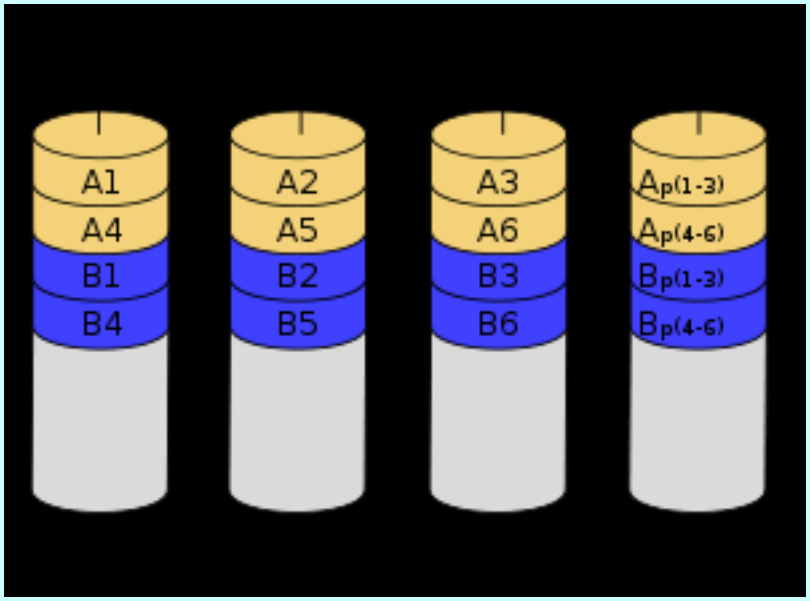


RAID 0 Stripping de bloques (hasta de 1 byte)



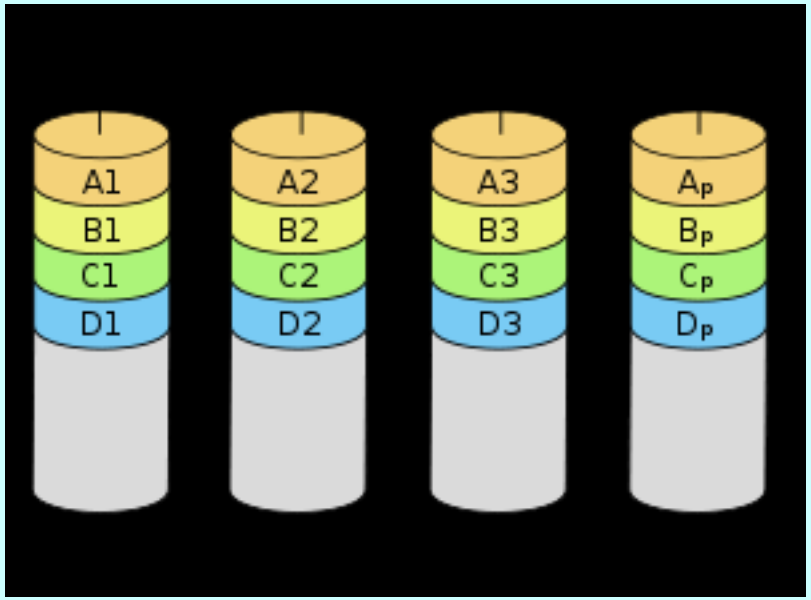
RAID 1 Mirroring

RAID2 Stripping de bits

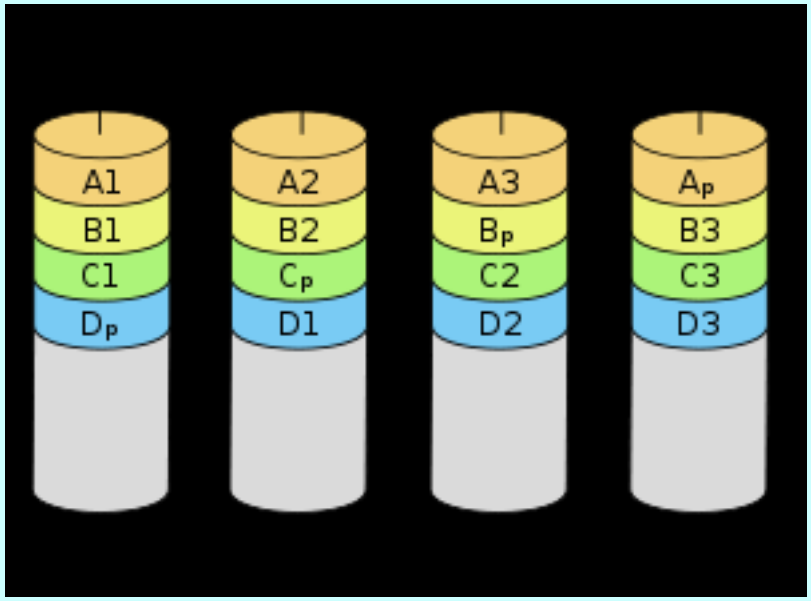


RAID 3 Byte Stripping con disco de paridad

Niveles de RAID

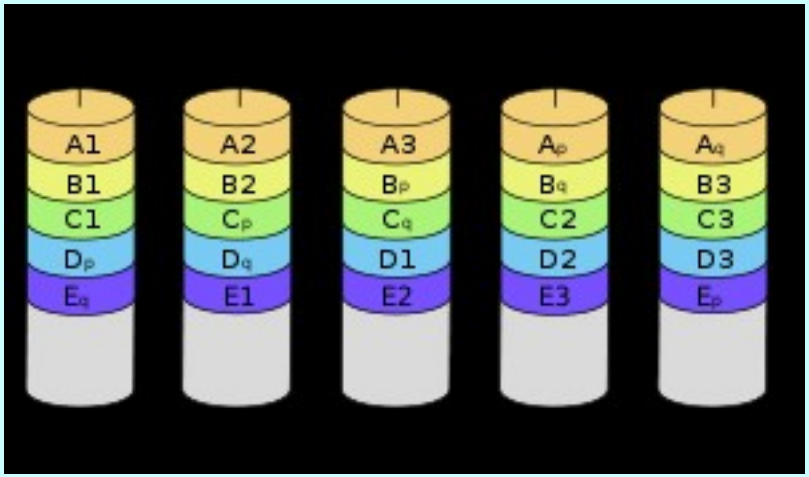


RAID 4 Block
Stripping con disco
de paridad

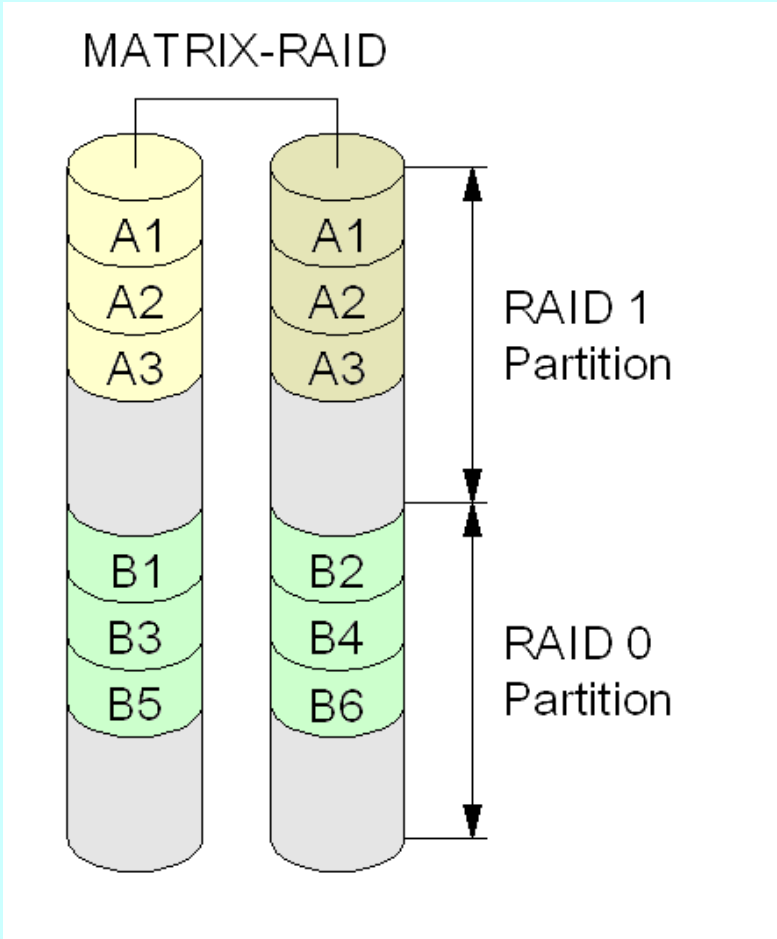


RAID 5 Block
Stripping con paridad
distribuida

Niveles de RAID



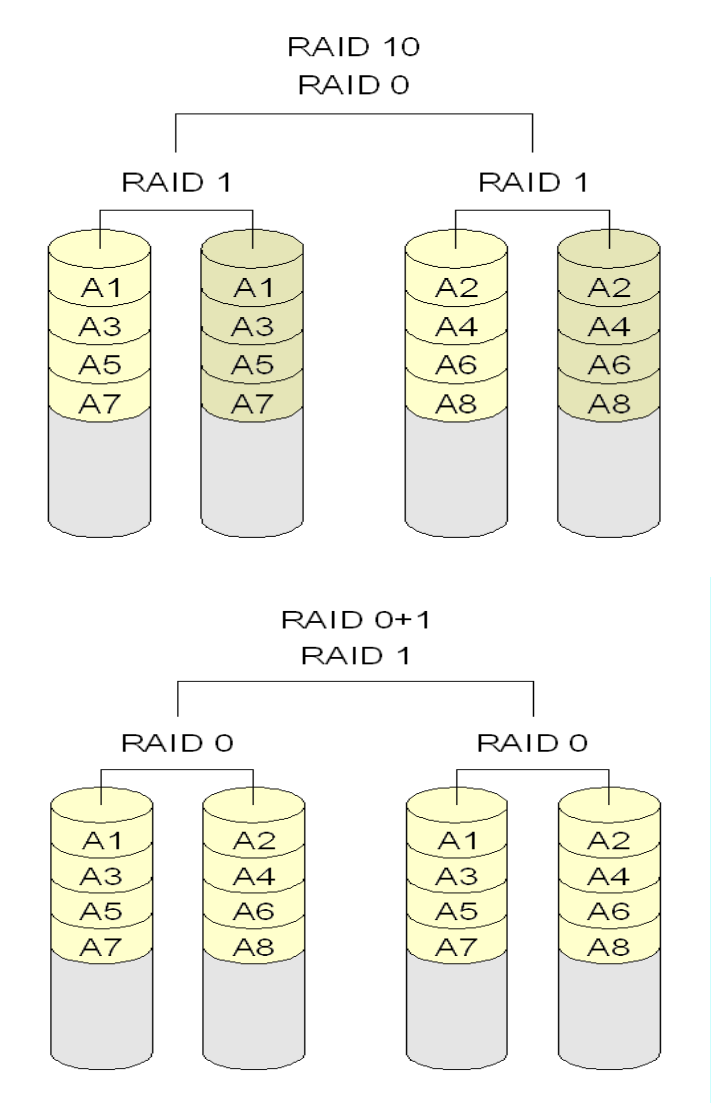
RAID 6 Block Stripping con doble paridad distribuida



Intel Matrix RAID

Raid Anidados

- Muchos de estos niveles están en software.
  - Raid 01 (ó 0+1)
  - Linux Raid 10 md.
    - Multiple Devices, creados a partir de uno o mas dispositivos independientes
  - Disponible a partir del Kernel 2.6



Software Raids

- El procesador debe usar su tiempo para las operaciones de RAID.
- En una capa entre el File System y el Device Driver.
- Grub lee RAID 1

Sistema Operativo	*Raids*
MAC OSX Server	RAID 0, 1, 1+0
Linux	Raid 0,1,2,3,4,5,6 y combinaciones
Windows Server	RAID 0, 1, 5

Hardware RAID

- Requiere de un controlador dedicado.
  - Debe tener un Back End hacia los discos ATA (PATA o EIDE), SATA, SCSI, Fibre Channel (que no necesariamente requiere fibra óptica) o SAS.
  - Un front end hacia el Host (usando un Host Adapter)
    - Que puede ser uno de los anteriores y ofrecer transparencia al acceso.
    - Algunos mas específicos como FICON,ESCON, iSCSI, HyperSCSI, ATA_over_Ethernet o InfiniBand

Disk Array Controllers

NEC-express
Sun Storage-Tek
EMC Clariion

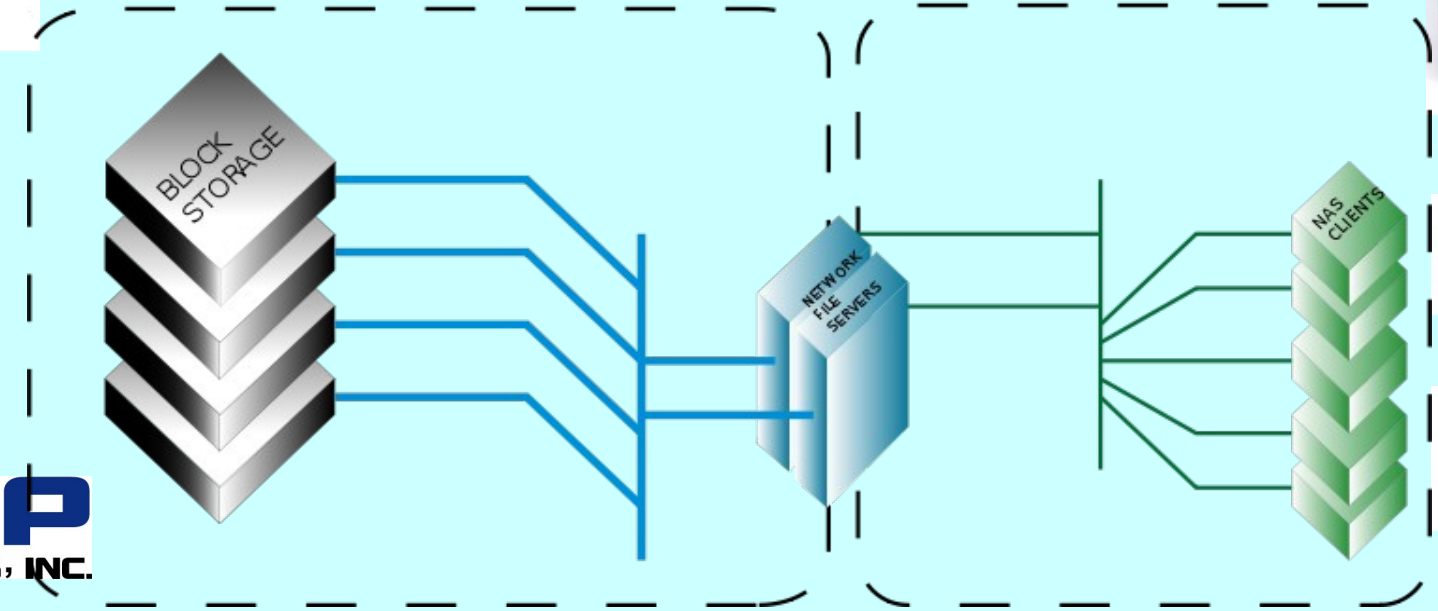
“Fake-Raid”

- Es un controlador de firmware que toma las funciones de raid durante el boot.
  - Una vez que el kernel de un SO está cargado, el control pasa al SO.
  - Se debe a que Windows no puede bootear de software RAID.
- Es un software raid y carga al procesador.
  - Con un controlador de múltiples canales ATA

NAS y SAN

- Network Attached Storage conecta un file-sytem remoto a una red, proveyendo el acceso a clientes heterogéneos.
- Storage Area Network conecta dispositivos remotos que el SO ve como locales (e implementa el file system).

NAS y SAN



iomega

NETGEAR
Everybody's connecting.

QNAP
QNAP SYSTEMS, INC.

PLEXTOR

DataDirect NETWORKS

Seagate

openfiler

NAS

TURNKEY LINUX

- Provee servicios basados en archivos.
- Generalmente es una versión reducida empotrada de algún Sistema Operativo.
  - Nexenta, FreeNAS, OpenFiler, TurnKey
  - Ofrecen SMB/CIFS, NFS o AFP.
  - Y acceso FTP,ssh, Web y WEBDAV.

SAN

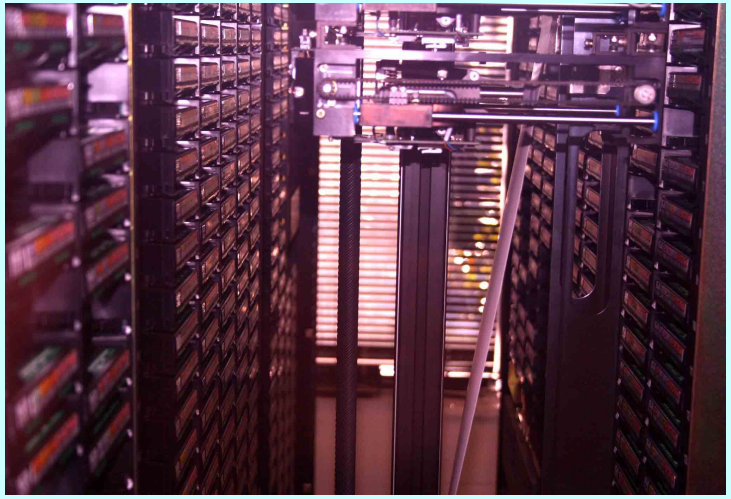
- Consolida las "islas de discos" con conexiones de red.
  - Pueden ser discos o RAIDs o alguna arquitectura no RAID
  - Usan protocolos como iSCSI, HyperSCSI, ATA_over_Ethernet o InfiniBand.
  - Requieren de un software de administración.
  - Algunas proveen capacidades RAID.

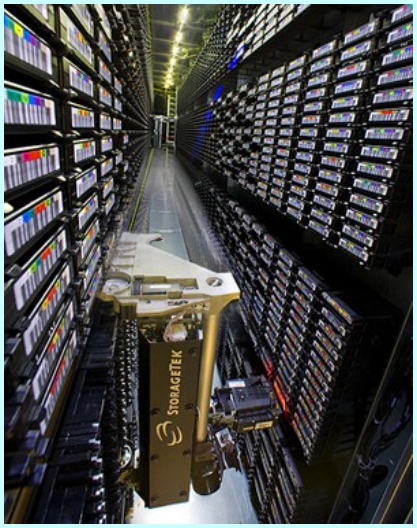
Almacenamiento de Red

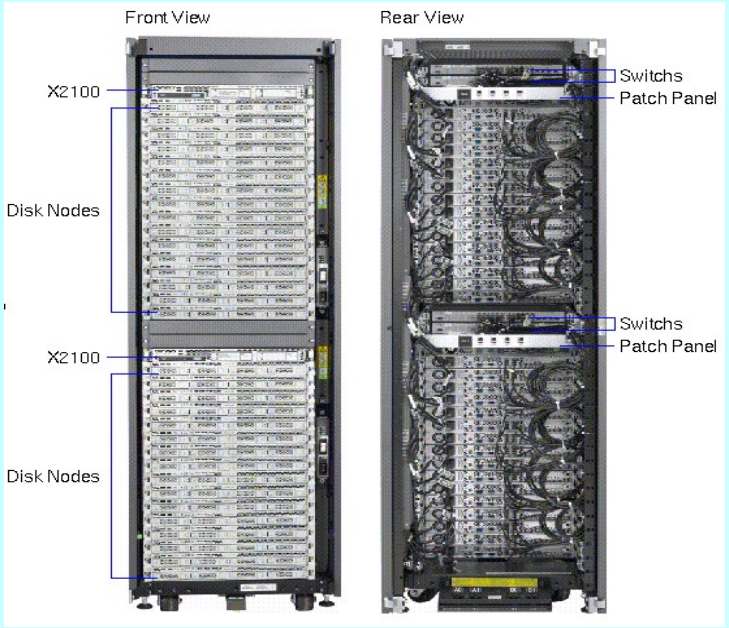
- Las plataformas existentes cubren un arco grande de prestaciones, tanto de bloques como de archivos.
  - Apple Xsan, IBM SVC, HP OpenVIew
  - Algunas usan Clustered File Systems o Shared Disk File Systems

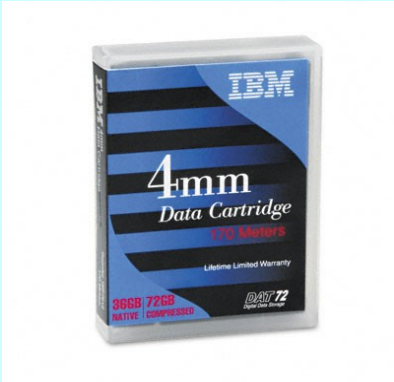
Otras Configuraciones de Almacenamiento

- Just a Bunch of Drives (JBOD) que permite expandir volúmenes.
- Massive array of idle disks (MAID) para aplicaciones 'Write Once, Read Occasionally' (WORO) con no mas del 25% de los discos simultáneamente activos.
- Configuraciones para Nearline Storage como Jukebox o CintoTecas






Front View
Rear View
X2100
Disk Nodes
X2100
Disk Nodes
Switchs
Patch Panel
Switchs
Patch Panel


IBM
4mm
Data Cartridge
Lifetime Limited Warranty
DAT 72


HP Magneto-Optical Disk
9.1GB