

Tema 4. Dispositivos de Almacenamiento

- *Juan Carlos Pichel*
- Enxeñería de Computadores
- Grao en Enxeñería Informática

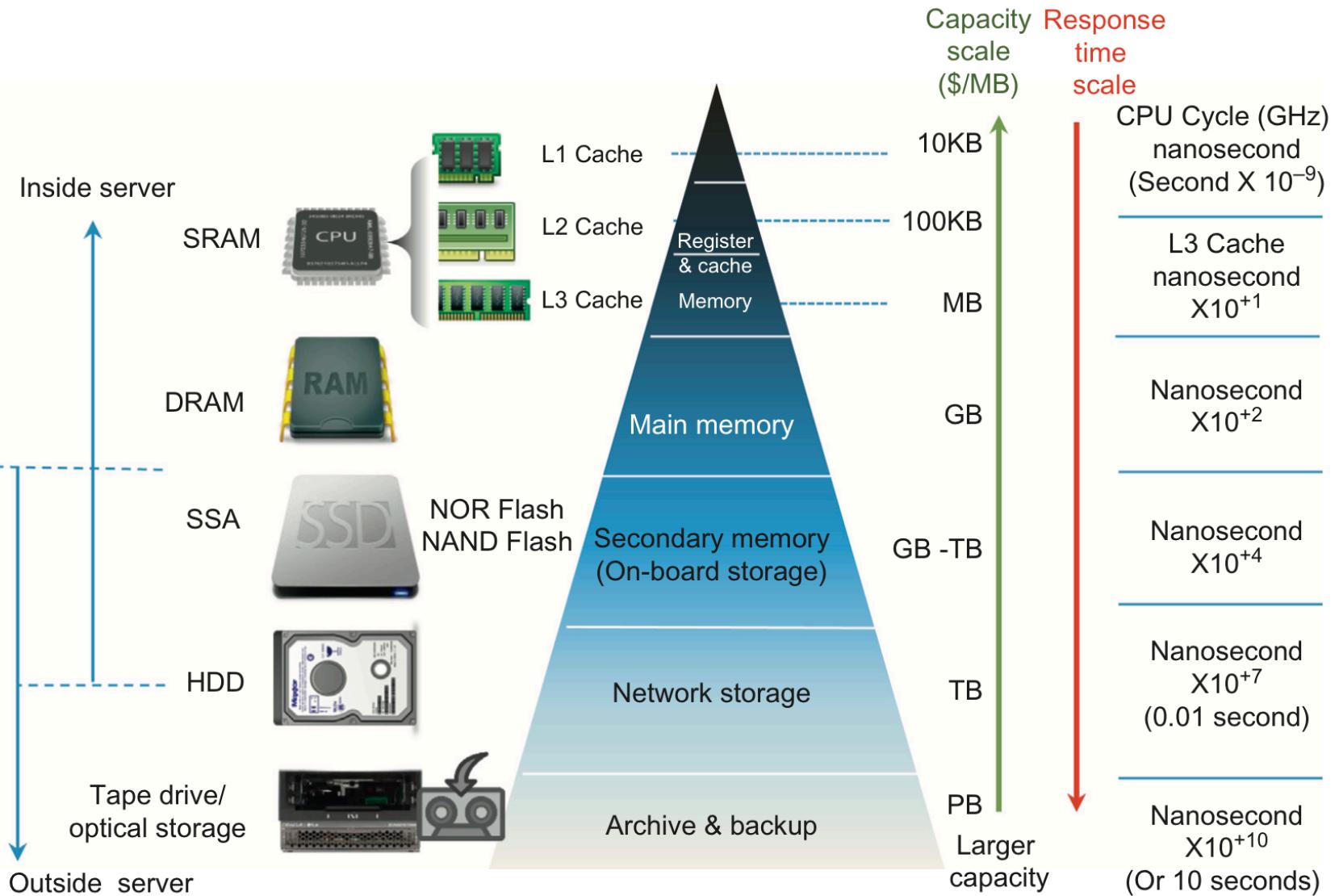
Almacenamiento

- Elemento central de muchos modernos CPDs
- La dificultad de la gestión supera a la del procesamiento
- Filosofía centrada en el almacenamiento

Algunas cifras significativas:

- La cantidad de datos que se almacena crece cada año el 40%, 50X en la próxima década (2016)
- Dar solución a las crecientes necesidades de almacenamiento es el principal problema de administración para el 79% de los profesionales (2016)
- La capacidad de discos instalada aumenta cada año el 60%
- En Unix y Linux, cada disco está ocupado al 30-50% (2009)
- En Windows, cada disco está ocupado al 20-40% (2009)

Almacenamiento



Discos magnéticos (HDD)

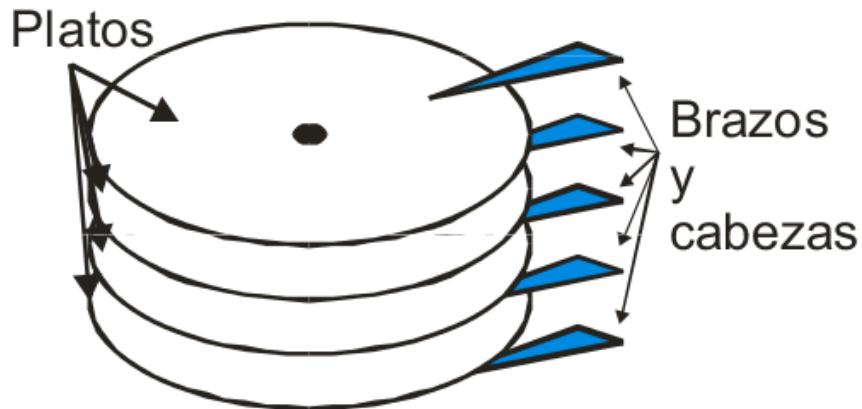
Son en la actualidad la **tecnología dominante**

- Densidad de almacenamiento creciente (1.4 Tbits/ pulgada², 2015 - Seagate)
- Continuas mejoras en la tecnología:
 - Heat-assisted magnetic recording (HAMR)
 - Bit-patterned media (BPM)
 - Shingled magnetic recording (SMR)

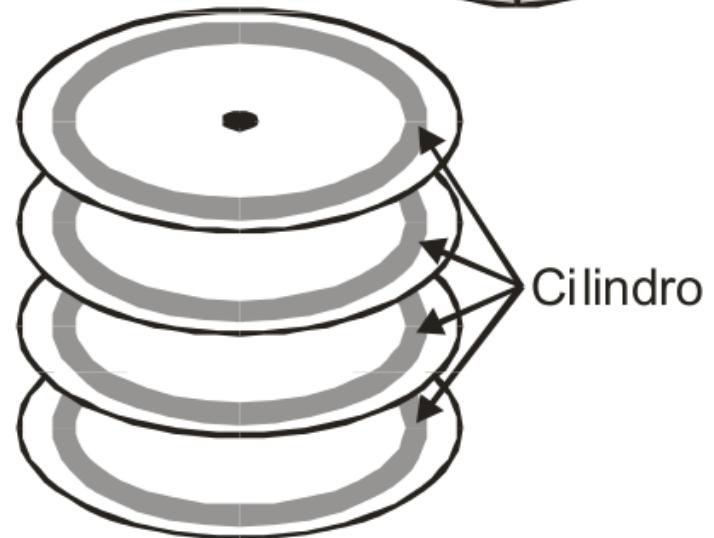
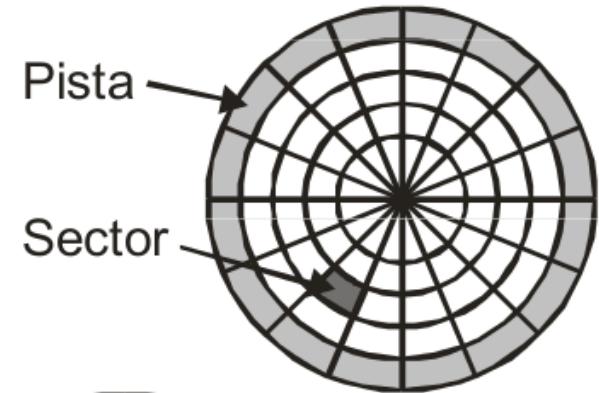
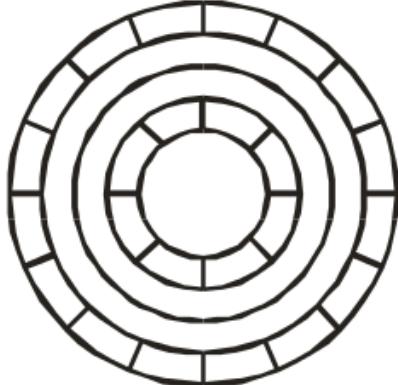
Otras alternativas: **discos de estado sólido (SSD)**

- Capacidad reducida
- Coste elevado
- Pero muy buenas prestaciones de velocidad y consumo
- Son utilizados en ciertas aplicaciones, se supone que coexistirán

Estructura física de un HDD



El número de sectores por pista
depende de la cercanía al centro



[Video](#)

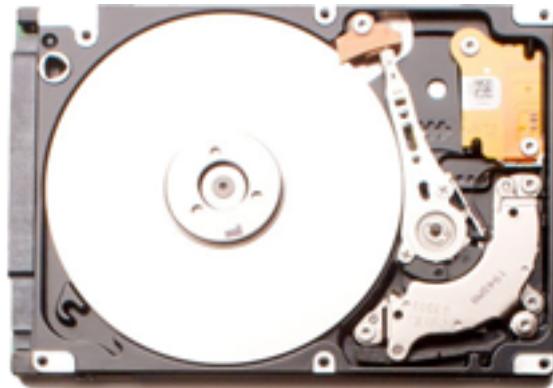
Prestaciones HDDs

Dependientes de:

- **Velocidad de acceso**, depende a su vez de:
 - Tiempo de acceso, necesario para situarnos sobre una pista dada
 - Latencia rotacional de la cabeza de lectura/escritura (acceso al sector)
 - Tasa de transferencia de datos (tiempo necesario para que los datos pasen por debajo de la cabeza)
- La **tasa de transferencia** depende de:
 - Velocidad de giro del disco
 - Características del interfaz (*una vez leídos los datos, enviarlos*):
 - Servidores: SATA, SCSI, SAS, FC – Doméstico: SATA, Wi-Fi, USB, Firewire...
- **La velocidad de giro del disco es el factor más importante** (entre 4200 y 15000 rpm):
 - En la mayoría de CPDs: 7200, 10000 y 15000 rpm

Formatos de HDDs

Anchura del disco (factor de forma): 3.5 y 2.5 pulgadas son los más comunes



Fiabilidad de los HDDs

- MTBF de discos SCSI: 1,500,000 horas
- MTBF de discos SATA: 600,000 horas
 - Versiones para servidores con fiabilidad comparable a SCSI
- Fallos anuales en discos en un CPD: 0.70% - 0.78%
- Corrección de fallos: Códigos CRC
- Predicción de fallos:
 - SMART en SATA
 - Mediante comandos en SCSI

Discos de estado sólido (SSD)

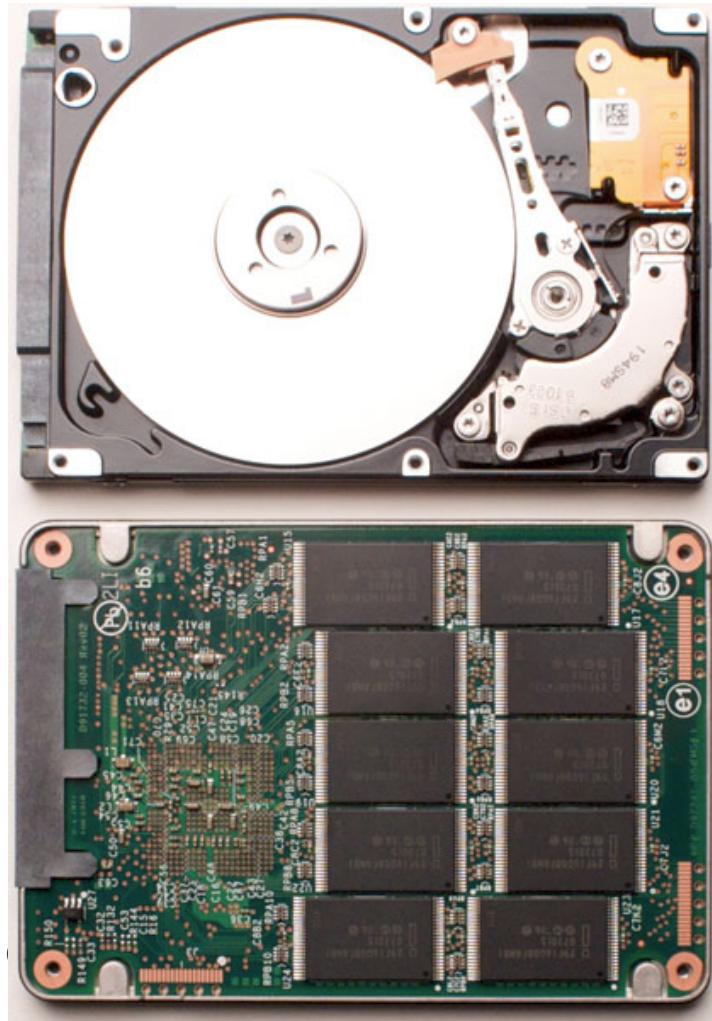
Reemplazo de los tradicionales

Ventajas

- Mayor velocidad de acceso
- Menor consumo
- Menor disipación de calor
- Menor peso
- Menor ruido
- Mayor MTBF (2M horas)

Desventajas

- Menor capacidad
- Mayor coste
- Mayor dificultad de recuperación en caso de fallo

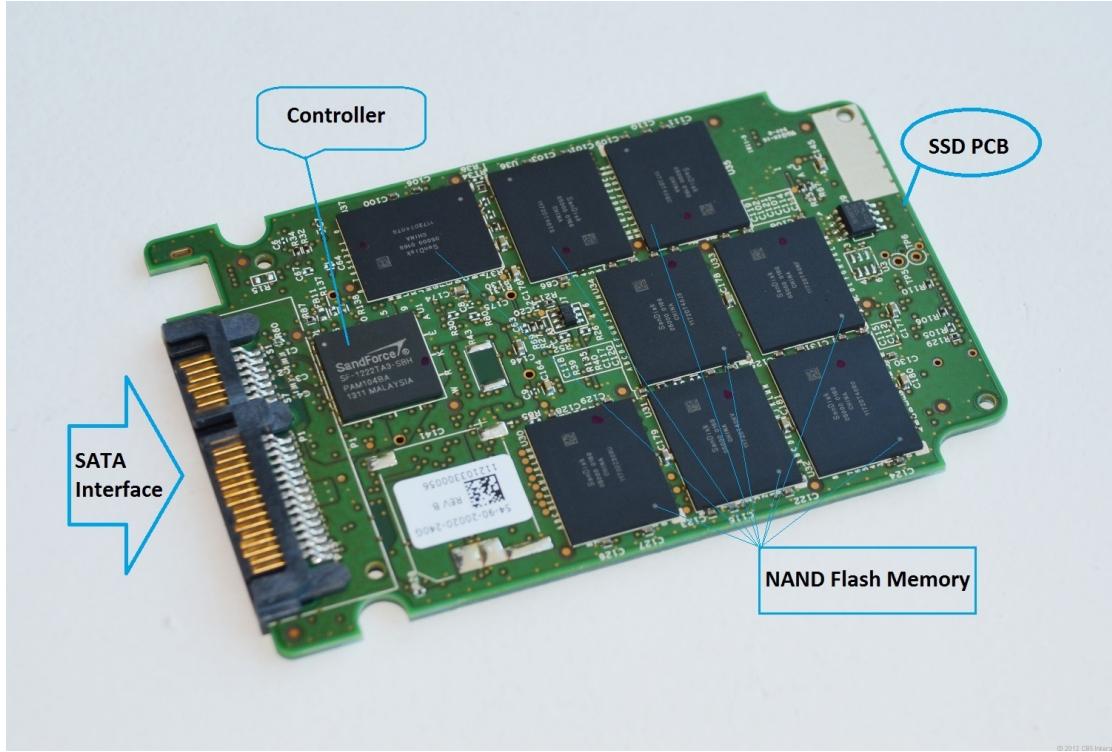


Discos de estado sólido (SSD)

Attributes	Solid State Drive (SSD)	Hard Disk Drive (HDD)
Physical size	2.5 inches	2.5 inches
Capacity	100/200/400 GB	146/300 GB
Media	SLC NAND	1–2 disks/2–4 heads
Interface	SAS 6 Gbps	SAS 6 Gbps
Avg. access/seek time	0.1 ms	3.0 ms(read)/2.7 ms (write)
Random read (IOPS)	90,000	385
Random write (IOPS)	16,000	325
Sequential read (64 K)	500 MB/s	200 MB/s
Sequential write (64 K)	250 MB/s	200 MB/s
Reliability MTBF	2.0 million hours	1.6 million hours
Nonrecoverable read errors	1 per 10^{17}	1 per 10^{16}
Voltage	5/12 V	5/12 V
Sleep/idle mode	<1 W	4.5 W
Operational mode	6.5 W	8.7 W
Noise	0 dB	3.3 dB
Weight	152 g	220 g
Warranty	5 years	5 years
Cost/GB	\$0.47–\$1/GB	\$0.04–\$0.075/GB

Discos de estado sólido (SSD)

Basados en tecnología NAND Flash



Algunos problemas:

- Write amplification
- Wear leveling

Discos de estado sólido (SSD)

Los datos no pueden sobreescibirse

	A	B	C
Block X	D	free	free
	free	free	free
	free	free	free
Block Y	free	free	free
	free	free	free
	free	free	free
	free	free	free

1. Four pages (A-D) are written to a block (X). Individual pages can be written at any time if they are currently free (erased).

	A	B	C
Block X	D	E	F
	G	H	A'
	B'	C'	D'
Block Y	free	free	free
	free	free	free
	free	free	free
	free	free	free

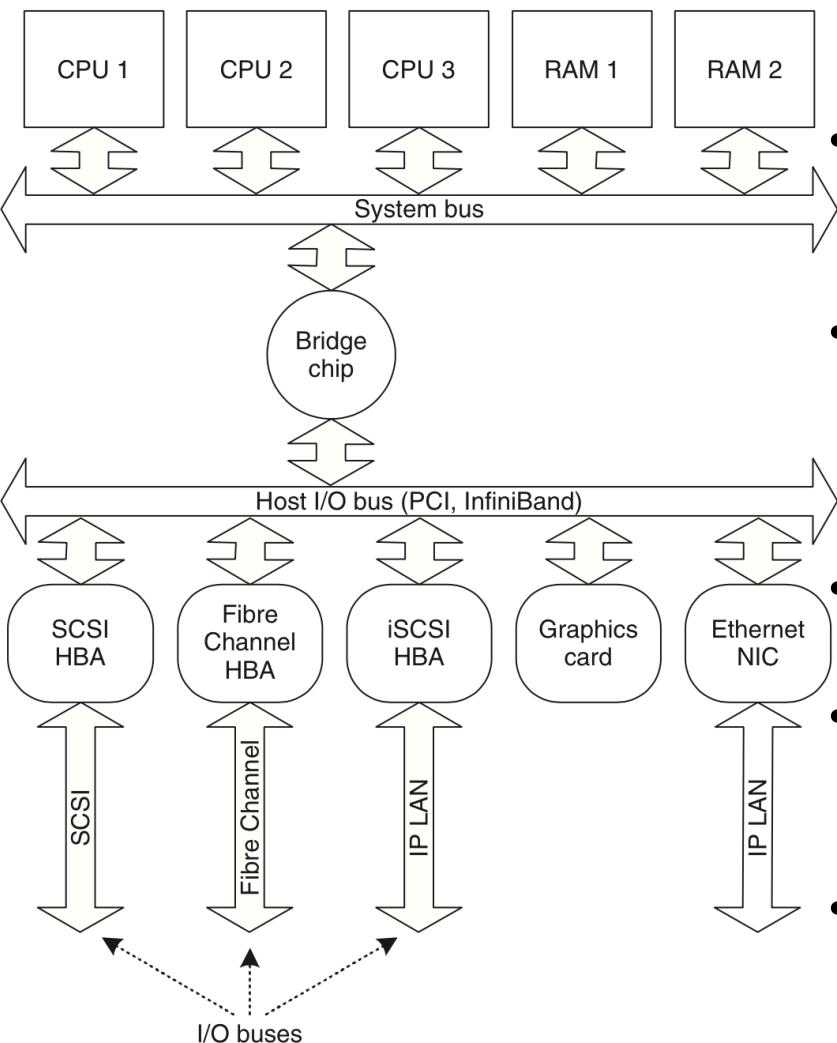
2. Four new pages (E-H) and four replacement pages (A'-D') are written to the block (X). The original A-D pages are now invalid (stale) data, but cannot be overwritten until the whole block is erased.

	free	free	free
Block X	free	free	free
	free	free	free
	free	free	free
Block Y	free	free	free
	free	E	F
	G	H	A'
	B'	C'	D'

3. In order to write to the pages with stale data (A-D) all good pages (E-H & A'-D') are read and written to a new block (Y) then the old block (X) is erased. This last step is *garbage collection*.

- Los bloques se estropean si usan muchas veces
- Se deben distribuir las escrituras a lo largo del disco

Camino físico de E/S entre CPU y el almacenamiento



- Datos desde RAM hasta los dispositivos de almacenamiento a través del **bus del sistema, bus de E/S del host y el bus de E/S**
 - **Bus del sistema:** conecta CPU y RAM, gran velocidad (frec. y BW), pocos dispositivos (corto)
 - Otros dispositivos no se pueden conectar a bus del sistema (limitaciones físicas) → **Bus de E/S del host** (PCI tecnología más extendida de este bus)
- **Bridge:** chip que conecta bus del sistema y el de E/S del host
- Comunicación periféricos: **Drivers** (programas en CPU y parte en ASICs contenidos en la placa (control. SCSI) o tarjetas de expansión (PCI))
- Disp. de almacenamiento se conectan al servidor a través del HBA (Host Bus Adapter) o de un controlador en la placa.
- Conexión de comunicación entre controlador y el periférico se llama **bus de E/S**

Interfaz de conexión y buses de E/S

Los discos son todos similares

- Los estándares de calidad son más altos para los discos de servidores

Existen varias opciones:

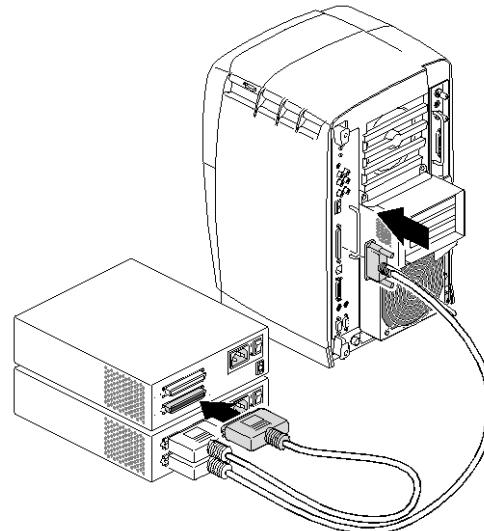
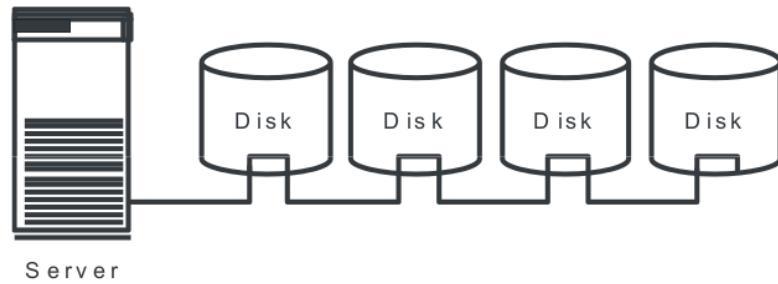
- PCs y estaciones de trabajo:
 - SATA
- Discos de gama alta para servidores usan **interfaces SCSI** que pueden ser de 3 tipos principalmente:
 - SCSI (Small Computer System Interface)
 - SAS (Serial Attached SCSI)
 - Fibre Channel

Para conexiones externas:

- USB
- Firewire
- eSATA
- ...

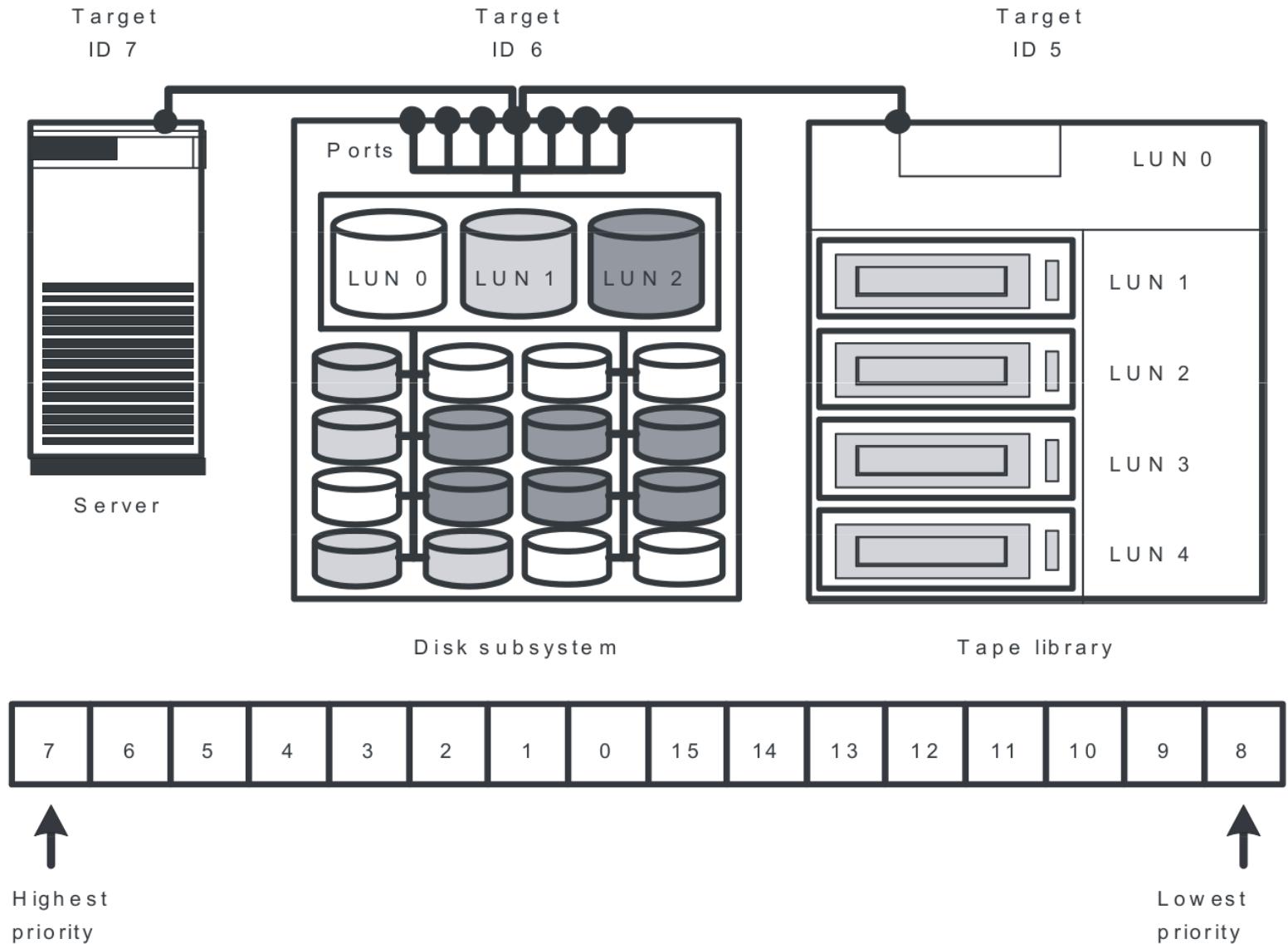
SCSI (Small Computer System Interface)

Interface	Tamaño del bus	Velocidad de reloj	Ancho de banda	Max. longitud del cable	Max. número de dispositivos
SCSI	8 bits	5 MHz	5 MB/s	6m	8
Fast SCSI	8 bits	10 MHz	10 MB/s	1.5-3m	8
Wide SCSI	16 bits	10 MHz	20 MB/s	1.5-3m	16
Ultra SCSI	8 bits	20 MHz	20 MB/s	1.5-3m	5 a 8
Ultra Wide SCSI	16 bits	20 MHz	40 MB/s	1.5-3m	5 a 8
Ultra2 SCSI	8 bits	40 MHz	40 MB/s	12m	8
Ultra2 Wide SCSI	16 bits	40 MHz	80 MB/s	12m	16
Ultra3 SCSI	16 bits	40 MHz DDR	160 MB/s	12m	16
Ultra-320 SCSI	16 bits	80 MHz DDR	320 MB/s	12m	16
Ultra-640 SCSI	16 bits	160 MHz DDR	640 MB/s	-	16



Protocolo SCSI: define como se comunican los dispositivos conectados al bus SCSI

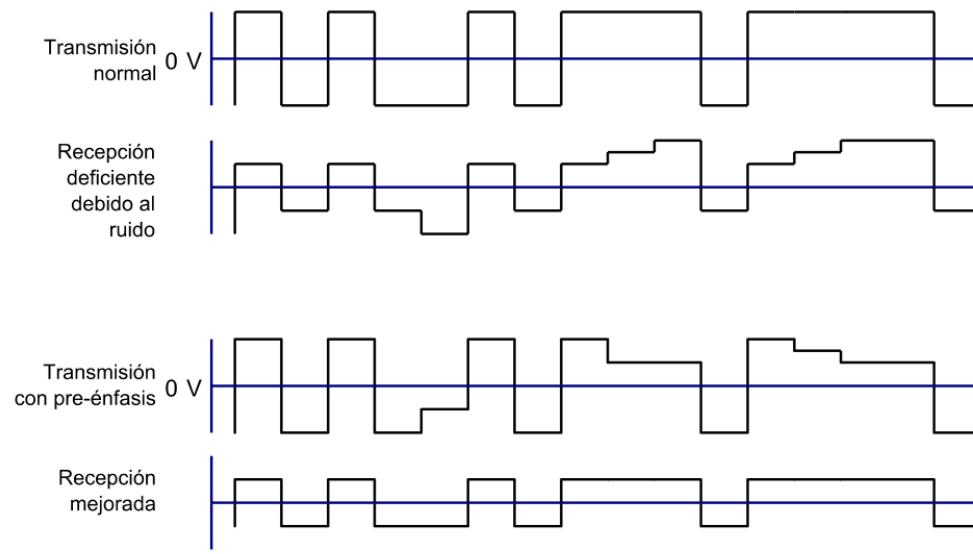
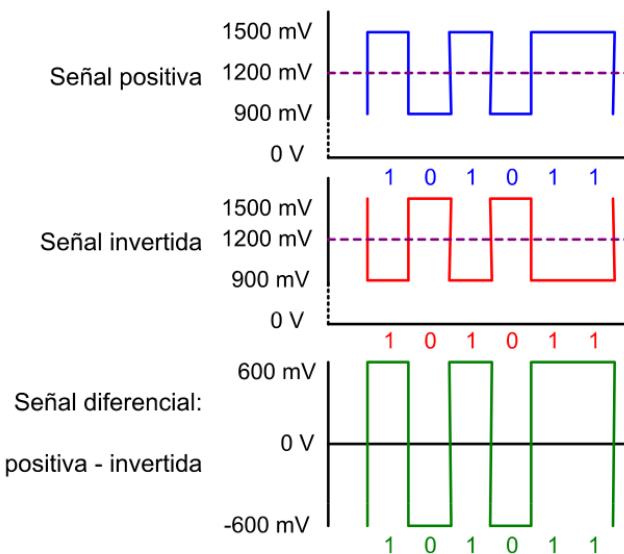
SCSI



Serial Attached SCSI (SAS)

Interface	Tamaño del bus	Velocidad de reloj	Ancho de banda	Max. longitud del cable	Max. número de dispositivos
SSA	1 bit	200 MHz	40 MB/s	25m	96
SSA 40	1 bit	400 MBHz	80 MB/s	25m	96
FC-AL 1Gb	1 bit	1 GHz	100 MB/s	0,5 - 3 km	127
FC-AL 2Gb	1 bit	2 GHz	200 MB/s		127
FC-AL 4Gb	1 bit	4 GHz	400 MB/s		127
iSCSI		Dependiente de la red			
SAS 3Gbit	1 bit	3 GHz	300 MB/s	6 - 8 m	16,256
SAS 6Gbit	1 bit	6 GHz	600 MB/s	10m	16,256

SAS-3: 12Gbit/s (1200 MB/s), SAS-4: 22.5 Gbit/s



Serial Attached SCSI (SAS)

Conexiones punto a punto full duplex

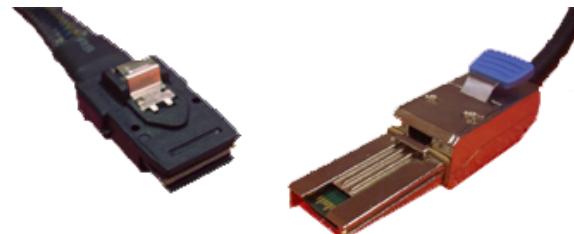
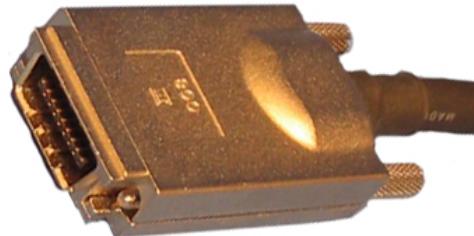
- No se comparte ancho de banda salvo si se utiliza un expansor

Hasta 16256 dispositivos por puerto y 65536 en total

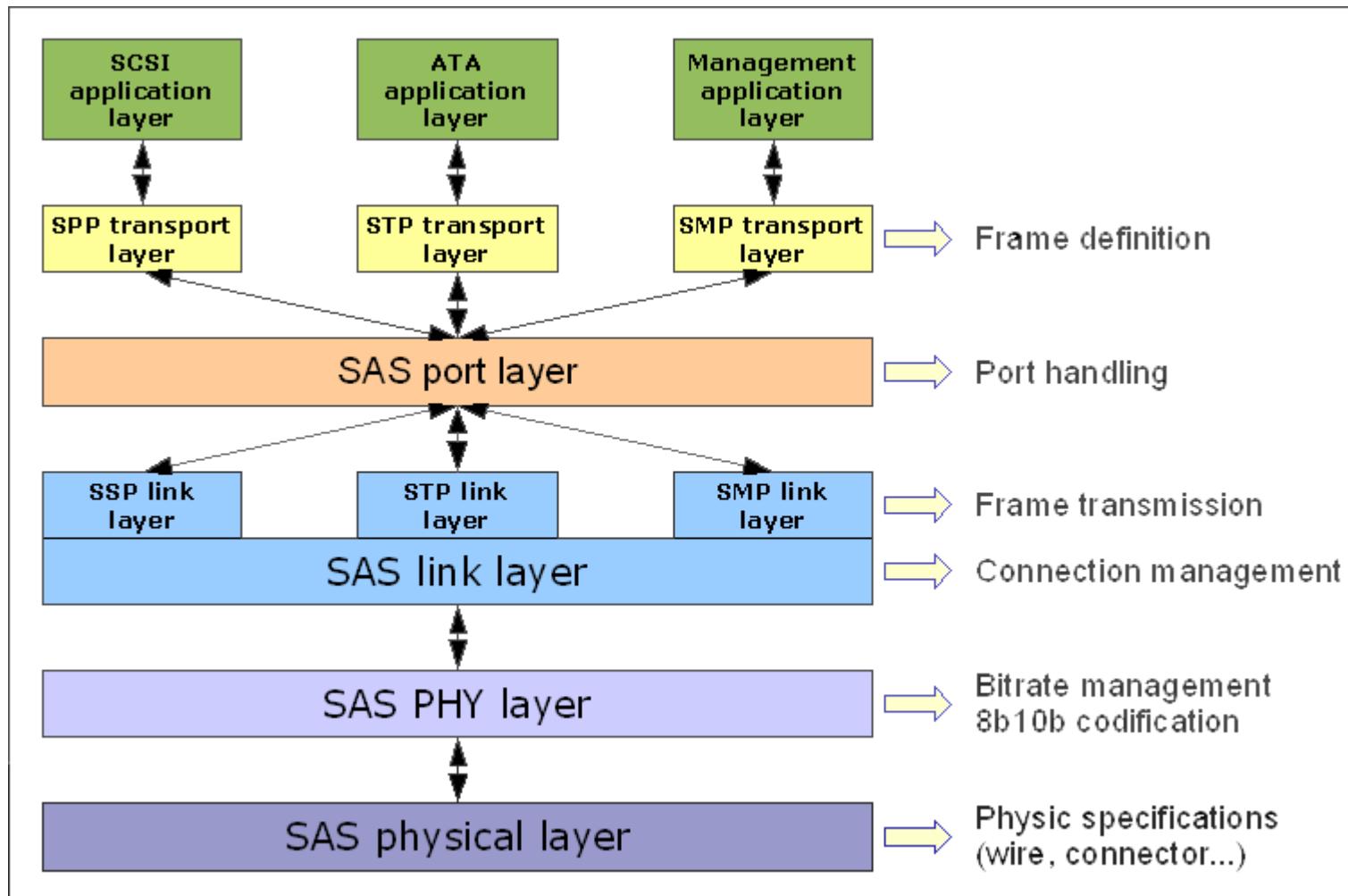
Utiliza los mismos comandos que SCSI

- También permite conectar dispositivos SATA

Se utilizan World Wide Names (WWN) en lugar de los ID SCSI



Arquitectura de SAS



Codificación 8b/10b

Utilizada por:

- PCI Express
- Firewire 800
- SATA
- SAS
- Fibre Channel
- SSA
- Gigabit Ethernet (no todos)
- InfiniBand
- DVI and HDMI
- DisplayPort Main Link
- HyperTransport
- USB 3.0
- y muchos otros

Cada dato de 8 bits se transmite como una palabra de 10 bits

La **redundancia** añadida permite:

- Detectar de errores
- Incluir palabras de control
- Mantener el número de unos y ceros equilibrado
- La señal de reloj se extrae de las palabras de información
 - No puede haber más de 5 bits iguales seguidos

Fibre Channel (FC)

- Los discos más rápidos y caros
- **Ventajas** del protocolo Fibre Channel:
 - Transmisión serie de alta velocidad a larga distancia (kms)
 - Baja tasa de errores de transmisión
 - Baja latencia
 - El protocolo FC (FCP) se implementa en hardware en el HBA para no cargar la CPU
- Permite construir redes de almacenamiento

Serial ATA

- Equipos domésticos y entornos semi-profesionales
 - Por su bajo precio tiene cierta penetración en entornos profesionales
- Bus serie como SAS
- También utiliza codificación 8b/10b
- Serial ATA 300 o SATA 3 Gbit/s
 - 300 MB/s
- Serial ATA 600 o SATA 6 Gbit/s
 - 600 MB/s

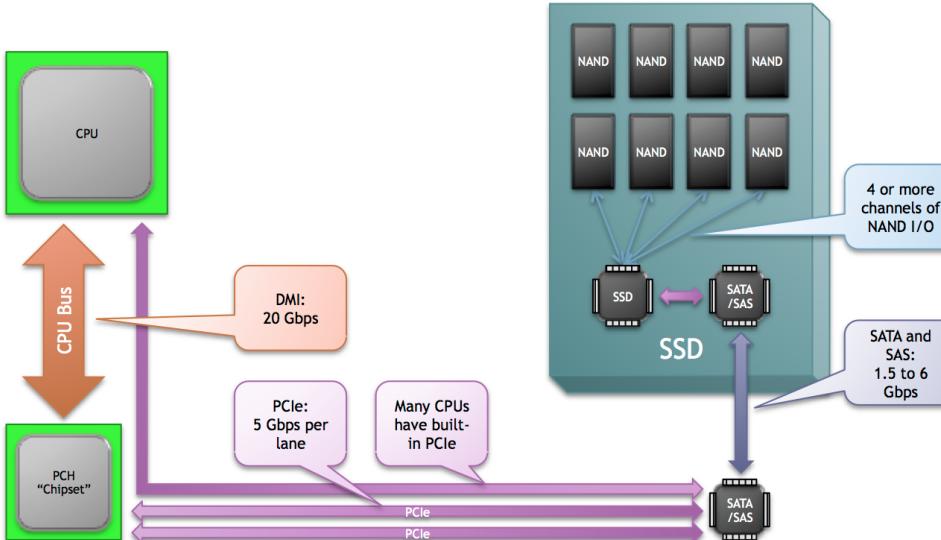


Comparación SAS - SATA

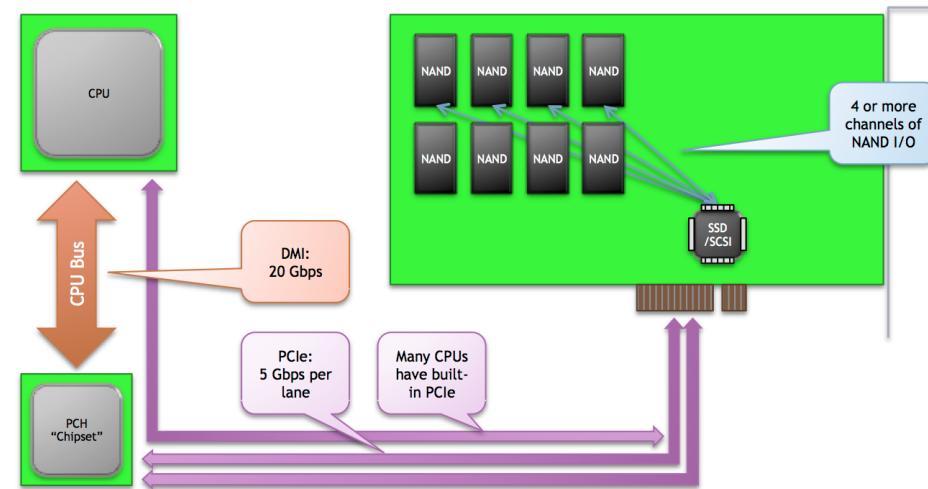
- SAS es full-duplex, SATA es half-duplex
- SAS utiliza WWN, SATA identifica los dispositivos por el puerto
- Gestión de secuencias de comandos
 - SAS: Tagged Command Queuing
 - SATA: Native Command Queuing
- SAS soporta unidades de CD/DVD, escáneres, impresoras, etc.
- SATA sólo discos y CD/DVD
- SAS permite **multipath** (más de una conexión al mismo dispositivo). Sólo algunas versiones de SATA lo hacen
- SAS es significativamente más caro que SATA.
- SAS detecta y corrige errores mejor que SATA (SMART)
- SAS utiliza voltajes más altos en sus señales.
 - Los cables SAS pueden extenderse hasta 10 metros, mientras que los SATA sólo 1 metro

Discos SSD PCIe

A Typical SSD SATA or SAS is the bottleneck



A PCIe SSD SSD controller or NAND is the bottleneck



Fuente: <http://blog.fosketts.net/2013/06/12/pcie-ssds-fast/>

Se conectan directamente al bus PCIe

- Evitan el overhead del controlador SATA o SAS
- Diferentes alternativas:
 - Tarjeta PCI Express
 - M.2 (factor de forma muy pequeño de disco que puede usar SATA y PCIe)

Discos SSD PCIe

Tarjeta PCI Express:



M.2 (factor de forma muy pequeño de disco que puede usar SATA y PCIe):



Organización de los discos

DAS (Direct-attached storage):

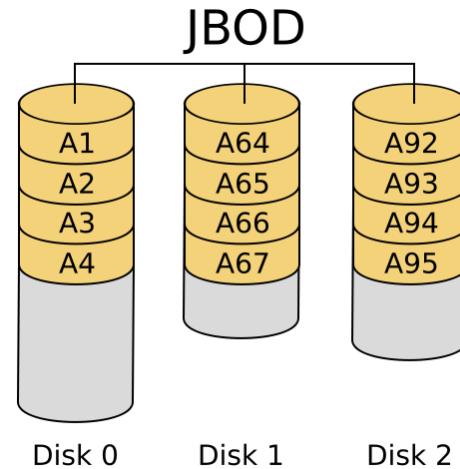
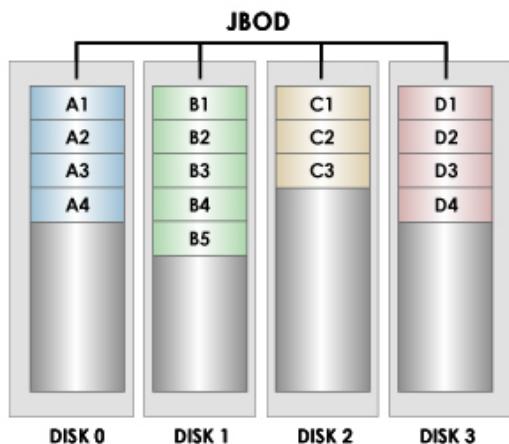
- Discos conectados directamente
- Conveniente para entornos sencillos

JBOD (Just a bunch of disks):

- Conjunto de discos están situados en el mismo chasis pero no están organizados de forma especial.

MAID (Massive array of inactive disks):

- Gran número de discos baratos a los que rara vez se accede



Organización de los discos - RAID

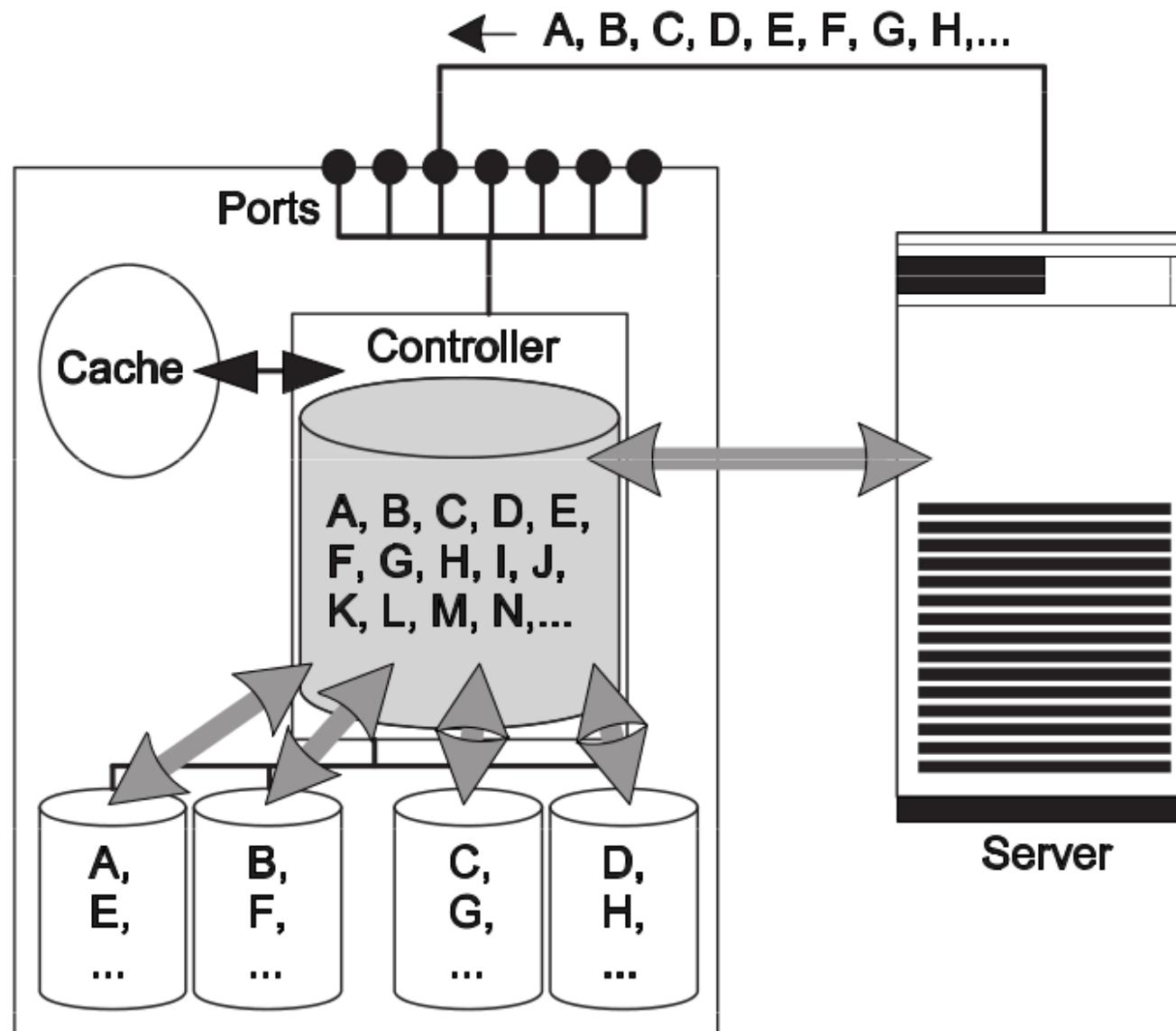
Redundant array of independent disks (RAID):

- Es la opción más eficiente para organizar varios discos
- Existen varias modalidades de RAID, llamados niveles

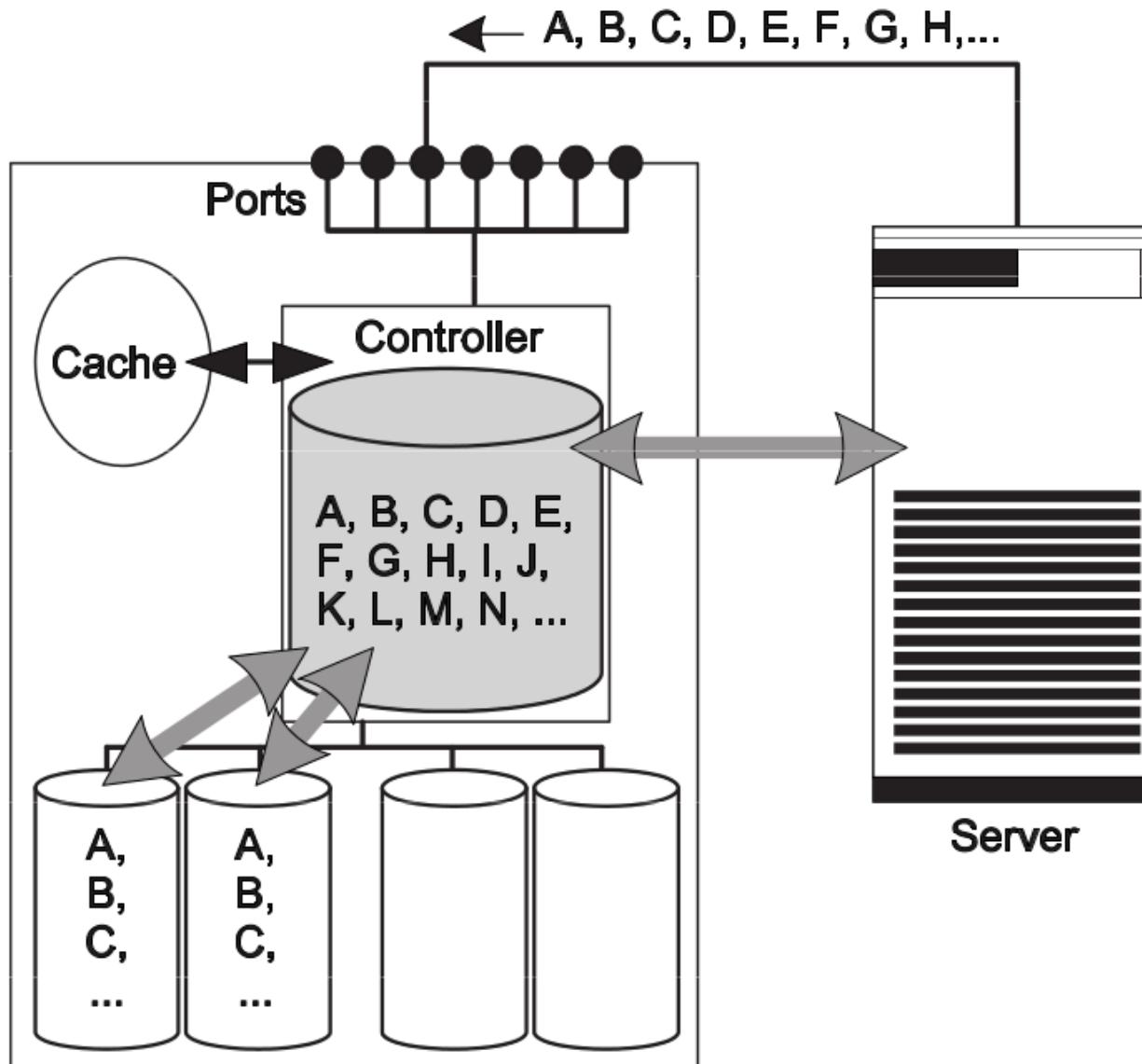
Niveles:

- 0 Striping: almacenamiento entrelazado
- 1 Mirroring: duplicado de datos
- 0+1 Mirrored stripes
- 10 Striped mirrors
- 3 y 4 Paridad localizada
- 5 Paridad distribuída
- 6 Paridad distribuída y duplicada

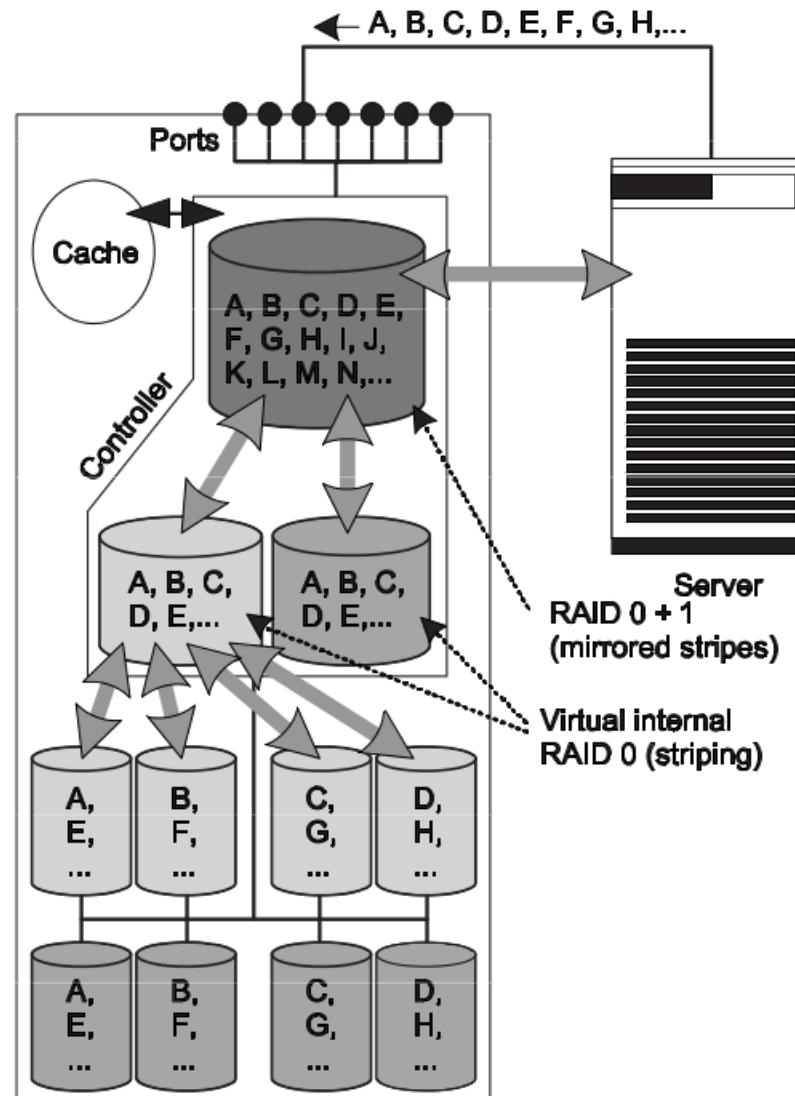
RAID 0



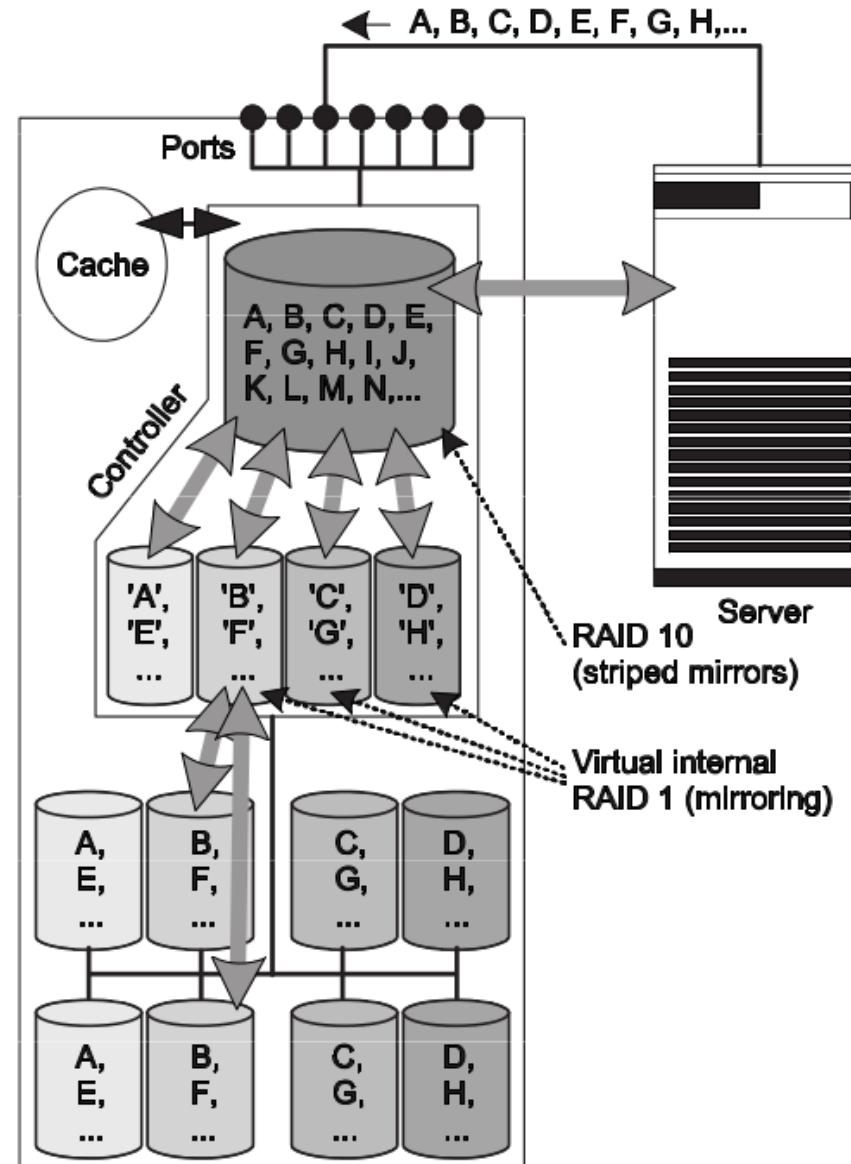
RAID 1



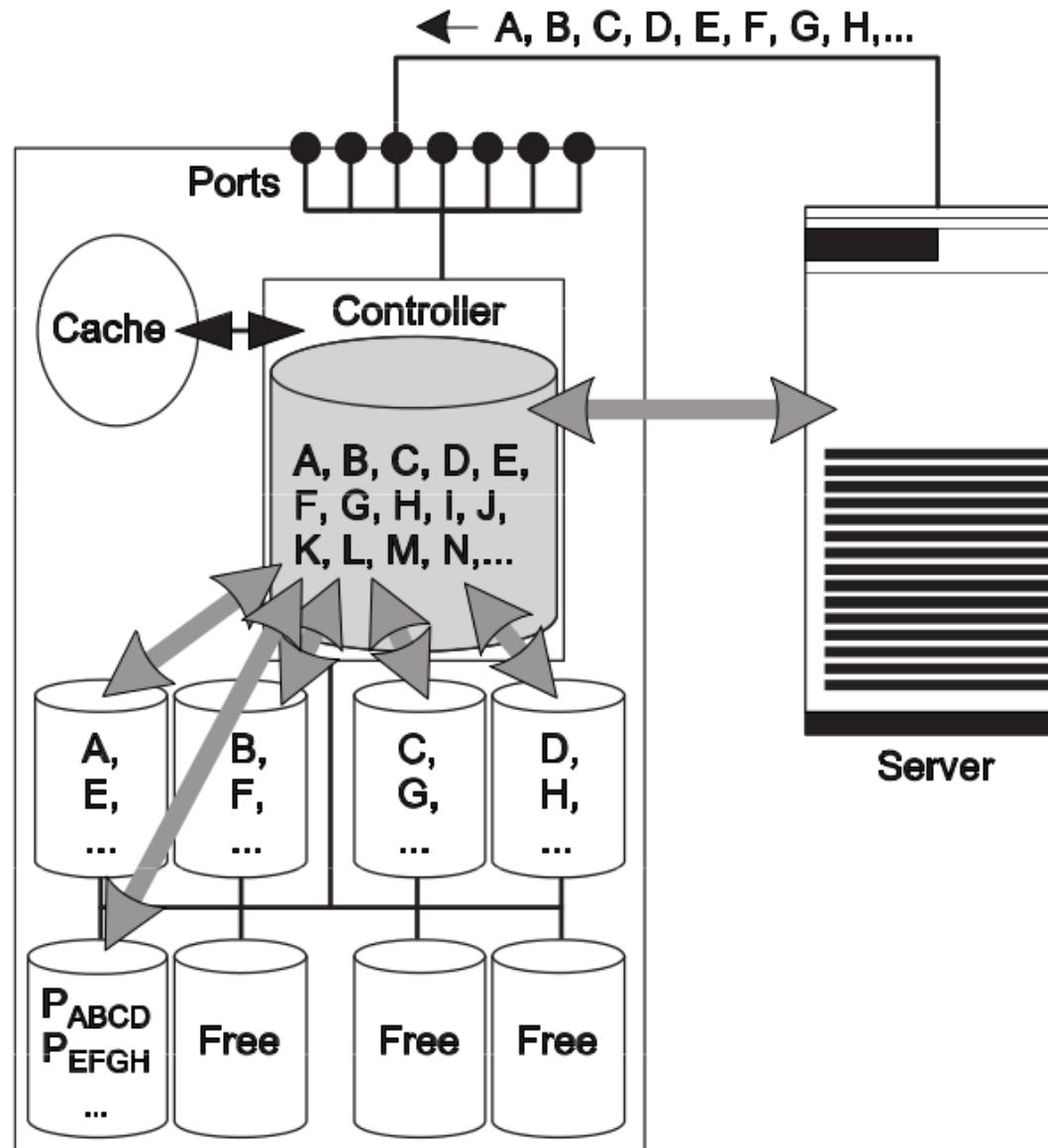
RAID 0+1



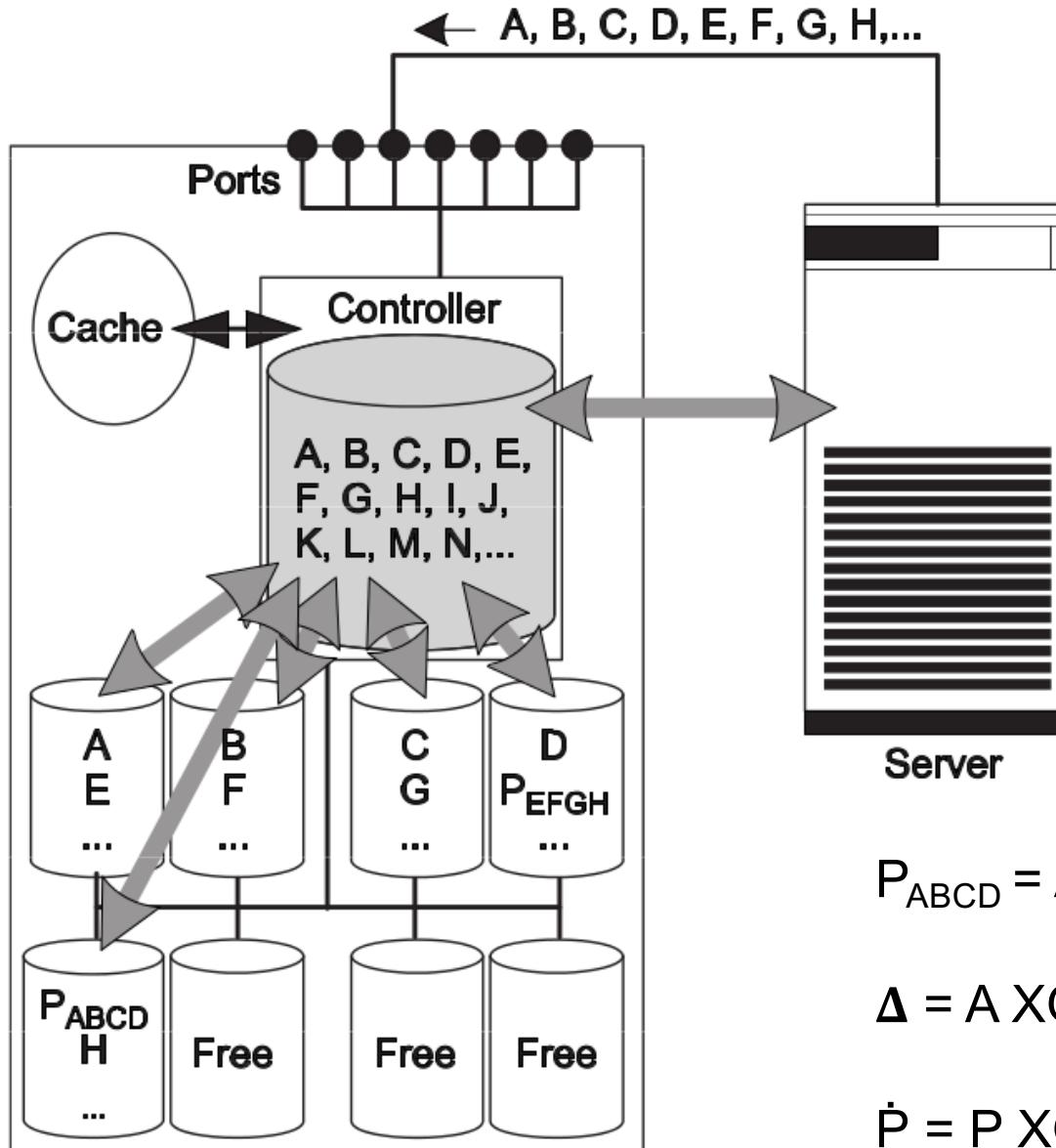
RAID 10



RAID 3 y 4



RAID 5



$$P_{ABCD} = A \text{ XOR } B \text{ XOR } C \text{ XOR } D$$

$$\Delta = A \text{ XOR } \bar{A}$$

$$\dot{P} = P \text{ XOR } \Delta$$

Resumen niveles RAID

RAID level	Fault-tolerance	Read performance	Write performance	Space requirement
RAID 0	None	Good	Very good	Minimal
RAID 1	High	Poor	Poor	High
RAID 10	Very high	Very good	Good	High
RAID 4	High	Good	Very very poor	Low
RAID 5	High	Good	Very poor	Low
RAID 6	Very high	Good	Very very poor	Low

RAID 0: Para maximizar la velocidad de escritura cuando la fiabilidad no es importante.

RAID 10: Para muchas escrituras a alta velocidad y gran robustez

- (logs de transacciones en bases de datos)

RAID 4 y 5: Seguros y baratos, pero bajo rendimiento en escritura

RAID 6: Para archivar datos.

- Mejora la escritura con caches de disco

Caches de Disco

- Acelerar operaciones de lectura y escritura
- Necesarias porque la velocidad del bus de I/O es mucho mayor que la velocidad a la que el disco puede leer o escribir
- **Para lectura:**
 - Se leen datos a la cache y se envían todos juntos, minimizando la ocupación del interfaz y el bus
 - Se aprovechan los posibles aciertos
- **Para escritura:**
 - Se almacenan los datos en la cache y se escriben a medida que es posible
 - Mejora la velocidad de escritura de los modos RAID 1, 4, 5 y 6

Caches en el RAID

- Las caches de los RAID funcionan a mayores de las cache de los discos
- **Cache de escritura**
 - Puede llegar a ser de GBs
 - Por seguridad, tiene una batería de apoyo y puede estar replicada
 - Las escrituras se almacenan en la cache donde estarán seguras aunque falle la electricidad
 - La escritura en disco se hará en cuanto sea posible y al ritmo que sea posible
 - La cache de escritura minimiza el impacto de los picos de escrituras seguidos de periodos sin ninguna escritura
 - También de las escrituras de muchos datos de forma secuencial
 - Esto es especialmente interesante par los modos de RAID más altos
- **Cache de lectura**
 - Analizan los patrones de acceso tratando de adivinar qué información debe precargar desde el disco duro
 - Como las lecturas son demasiado aleatorias, se considera que es un buen resultado acertar en el 40%