

upna

Universidad Pública de Navarra  
Nafarroako Unibertsitate Publikoa

**Redes de Nueva Generación**  
*Área de Ingeniería Telemática*

# Almacenamiento

upna

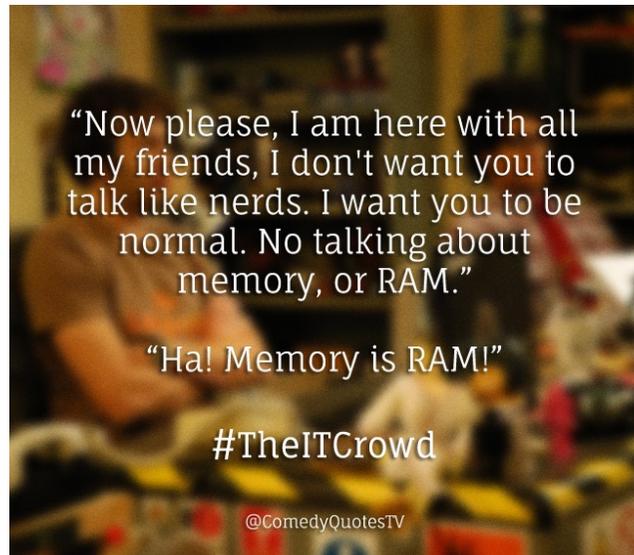
Universidad Pública de Navarra  
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación  
*Área de Ingeniería Telemática*

# Sistemas de almacenamiento

# Sistemas de almacenamiento

- Almacenamiento primario
  - Memoria RAM; volátil
  - Accesible directamente por la CPU
  - Acceso aleatorio
  - Pequeña capacidad y bajos tiempos de acceso



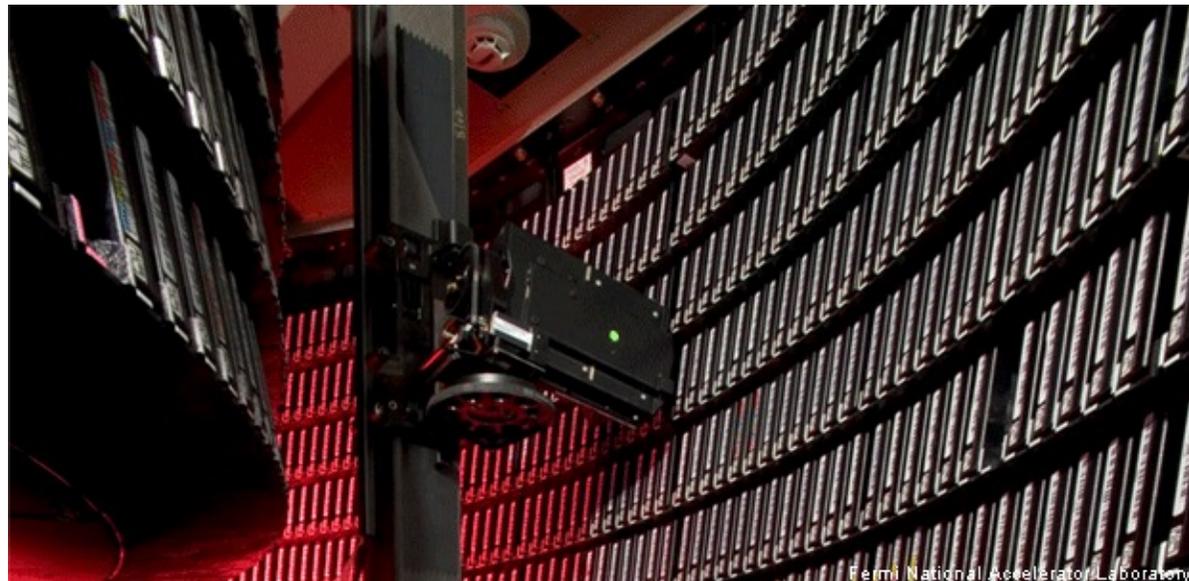
# Sistemas de almacenamiento

- Almacenamiento primario
- Almacenamiento secundario
  - No volátil
  - No es accesible directamente por la CPU
  - Requiere dispositivos de entrada salida (I/O)
  - Mayor capacidad y mayores tiempos de acceso
  - Acceso aleatorio
  - Discos duros



# Sistemas de almacenamiento

- Almacenamiento primario
- Almacenamiento secundario
- Almacenamiento terciario
  - Sistemas de almacenamiento removibles
  - Tiempos de acceso aún mayores pero coste por GB menor
  - Empleados para almacenamiento “*long term*”
  - Cintas (desde los 50s) de acceso secuencial
  - Pueden almacenar centenares de petabytes (y hasta exabytes)



upna

Universidad Pública de Navarra  
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación  
*Área de Ingeniería Telemática*

# Sistemas de almacenamiento

upna

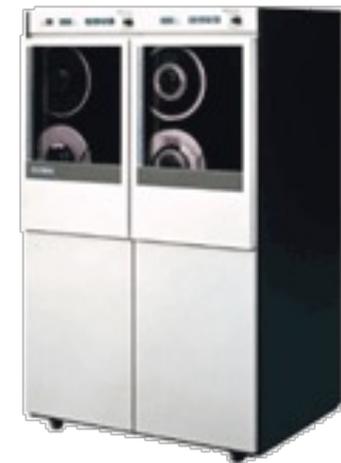
Universidad Pública de Navarra  
Nafarroako Unibertsitate Publikoa

**Redes de Nueva Generación**  
*Área de Ingeniería Telemática*

# El disco duro

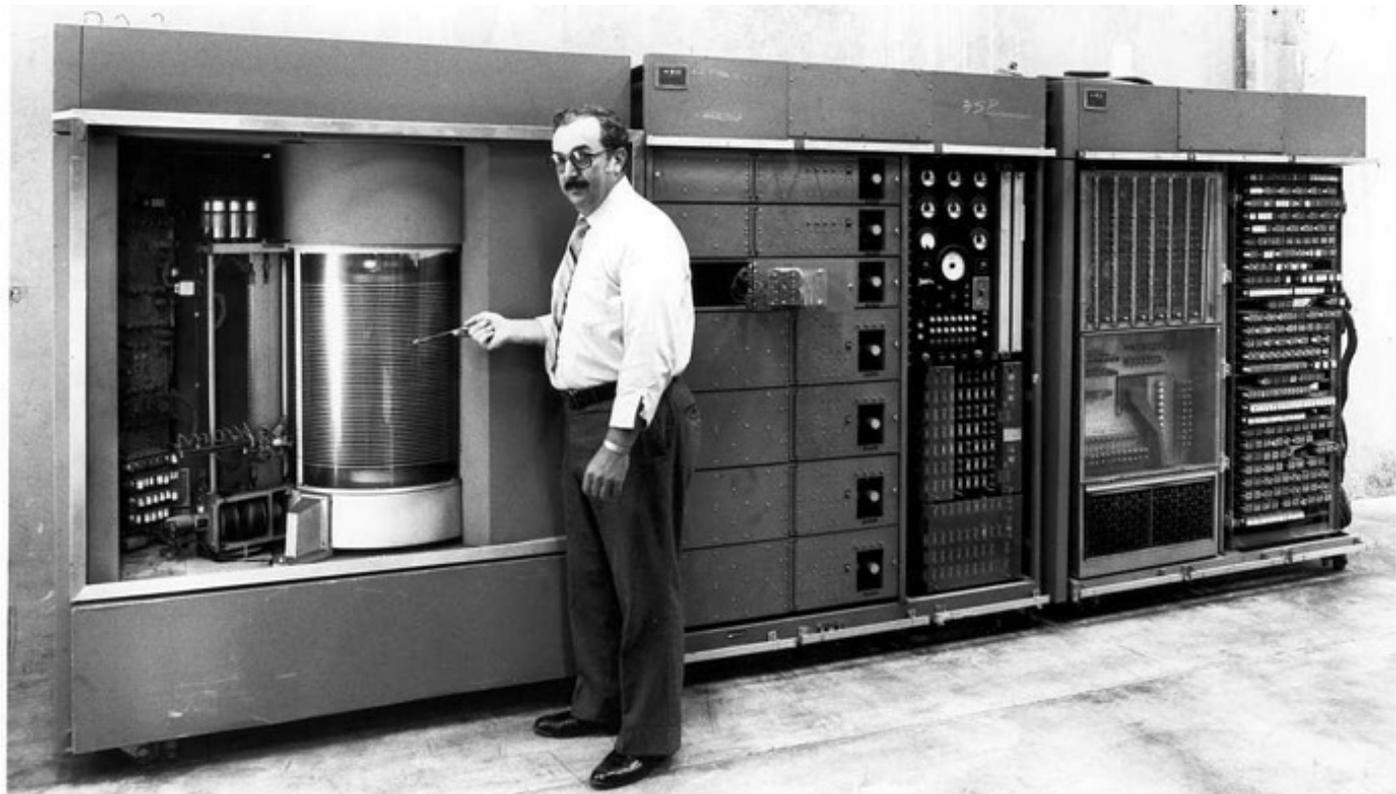
# Almacenamiento secundario

- Mainframes empleaban cintas magnéticas



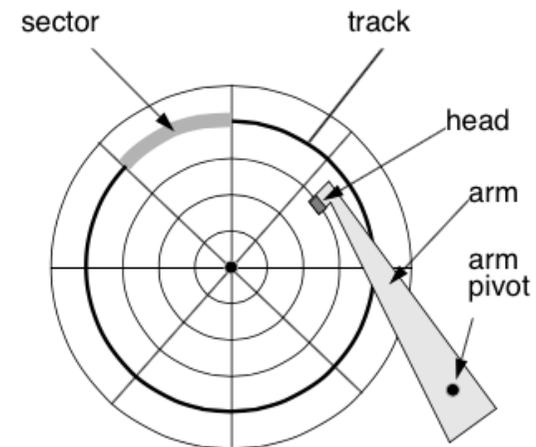
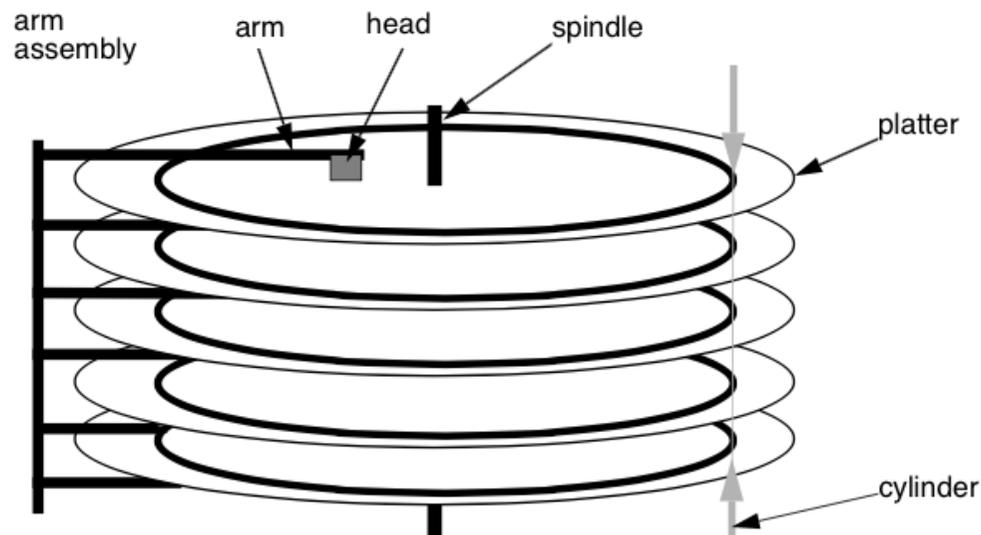
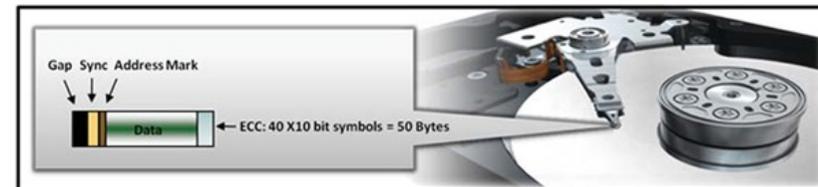
# Almacenamiento secundario

- A mediados de los 50 surge el disco duro o “*hard disk drive*”
- Sigue siendo magnético pero en lugar de una cinta son platos



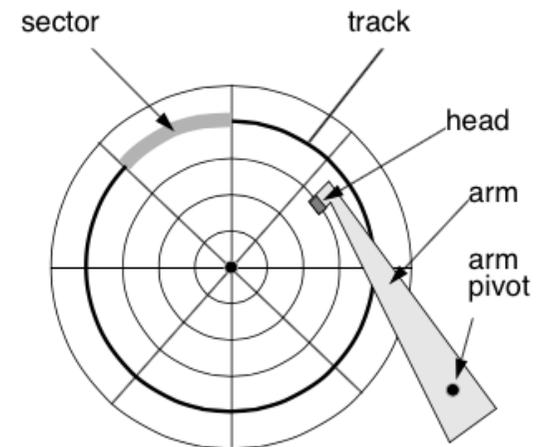
# Arquitectura del disco

- Platos (*platters*): material magnético
- Pistas (*tracks*): trayectoria circular
- Sectores: unidad mínima direccionable, todos del mismo tamaño, tradicionalmente 512 Bytes (hay ahora unidades con 4 KiB)
- Cilindros: pila vertical de pistas
- El tamaño del disco (en pulgadas) condiciona su capacidad y consumo
- Brazo y cabeza de lectura/escritura (por cada plato)
- Eje de rotación (*spindle*)



# Rotación del disco

- Todos los platos rotan al unísono
- Típicas velocidades de rotación: 7.200 rpm, 10.000 rpm, 15.000 rpm
- La cabeza debe avanzar hasta la pista donde están los datos
- Debe esperar a que el plato rote hasta que el sector que busca se encuentre debajo
- Entonces podrá leer o escribir (no a la vez)
- Lee del cilindro, así que cuando termina la pista pasa a la del mismo cilindro en otro disco
- Esas 4 operaciones llevan tiempo (posicionarse, esperar a que gire, leer/escribir y opcionalmente cambiar de plato o de pista)



# Tiempos básicos

- “*Seek time*”: Tiempo necesario para colocar la cabeza lectora en la pista deseada
- “*Rotational latency*”: Tiempo de espera a que el sector deseado alcance la cabeza
- “*Transfer time*”: Tiempo para transferir los datos del/al sector
- “*Bus transfer time*”: el protocolo del bus (SCSI, SATA) añade mensajes (handshakes)



(rpm)	Avg. rot. (ms)
5400	5.5
7200	4.2
10000	3
15000	2



# Discos “internos”

- En caso de fallos hardware los discos internos complican y enlentecen la reparación
- En el entorno servidor lo más común es que sean “cambiables en caliente” (*hot swappable*)
- En el caso del servidor el disco interno suele mantener el sistema operativo y caches
- Los datos estarán en discos externos



# Cabinas de discos

- Las cabinas de discos (*disk array*) pueden incluir una *controladora*
- Puede estar integrada con la cabina o con el servidor
- La cabina controla los discos y ofrece algún interfaz de acceso para el servidor (o los servidores)
- Los servidores podrán acceder a volúmenes lógicos creados en esos discos
- La controladora contará con una cache (RAM o flash)
- Puede contar con fuentes de alimentación redundantes



upna

Universidad Pública de Navarra  
Nafarroako Unibertsitate Publikoa

**Redes de Nueva Generación**  
*Área de Ingeniería Telemática*

# El disco duro

upna

Universidad Pública de Navarra  
Nafarroako Unibertsitate Publikoa

**Redes de Nueva Generación**  
*Área de Ingeniería Telemática*

# RAID

# RAID

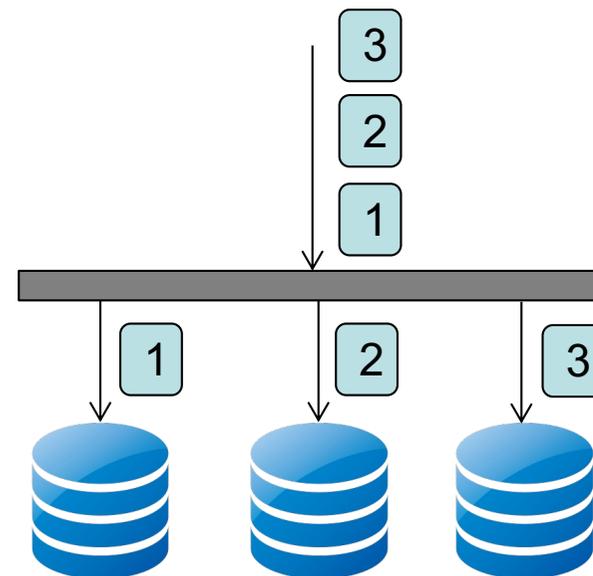
- *Redundant Array of Independent (Inexpensive) Disks*
- Varios discos que de cara al usuario (el servidor) se comportan como un solo volumen
- Los diferentes tipos de RAID se denominan mediante un “nivel”
- RAID level 0, RAID level 1, etc
- En comparación con *Just a Bunch Of Disks* (JBOD)



# RAID levels

## RAID 0

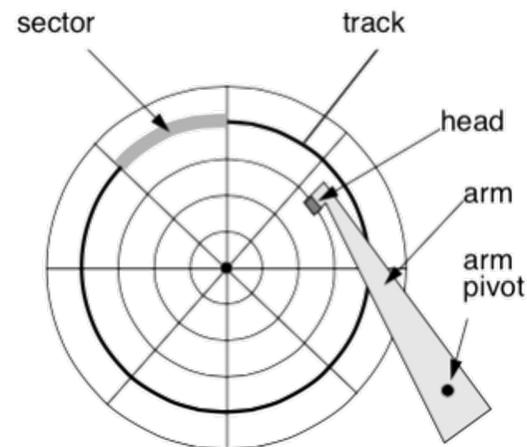
- *Disk stripping*
- Se reparten los datos entre varios discos
- Esto permite mayores velocidades de transferencia
- No hay redundancia
- Un fallo en un disco es irre recuperable
- Cualquier número de discos (mayor que 1)



# RAID levels

## RAID 0

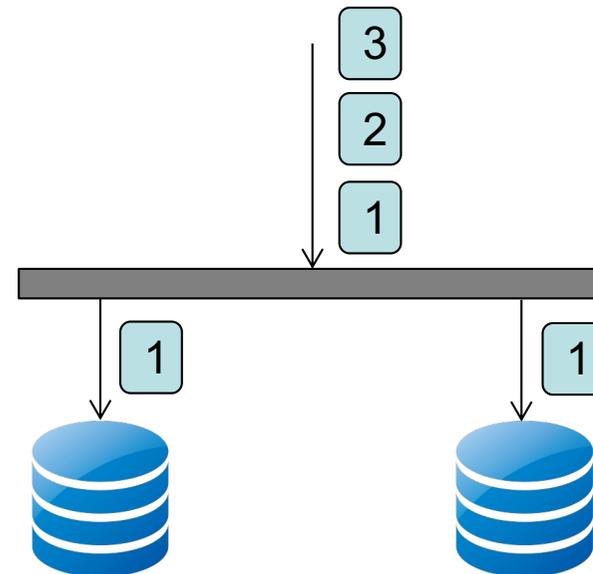
- *Disk stripping*
- Se reparten los datos entre varios discos
- Esto permite mayores velocidades de transferencia
- No hay redundancia
- Un fallo en un disco es irre recuperable
- Cualquier número de discos (mayor que 1)
  
- *Slicing*
  - La velocidad lineal es superior en las pistas exteriores
  - Se puede aumentar la velocidad creando el RAID empleando solo esos sectores



# RAID levels

## RAID 1

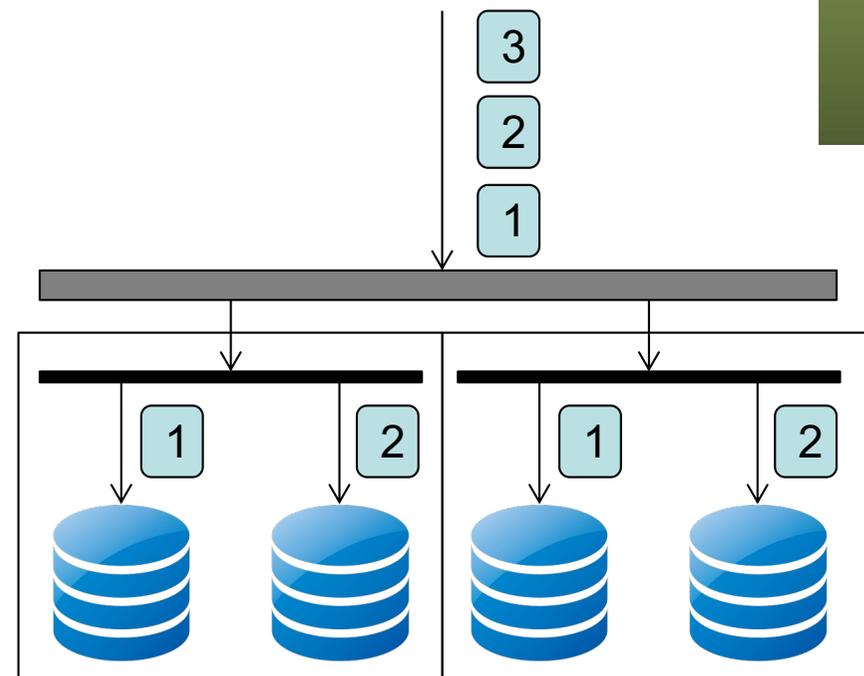
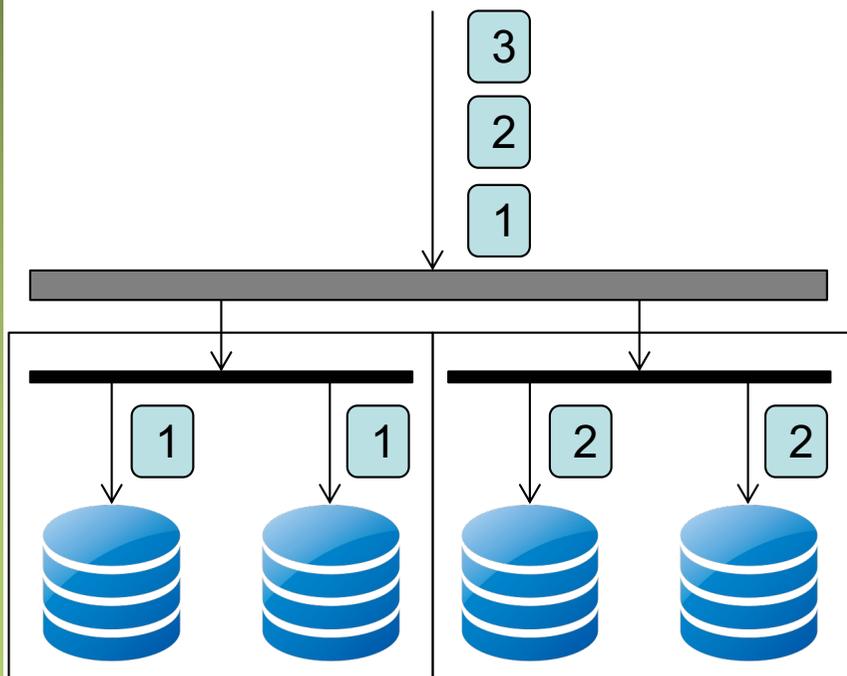
- *Mirroring*
- Los datos se replican
- Requiere al menos 2 discos
- No mejora la velocidad pero sí da protección
- Ante un fallo en un disco el RAID puede seguir funcionando
- Se puede sustituir el disco defectuoso y la controladora reconstruye el *mirror*
- La reconstrucción reduce el rendimiento del disco



# RAID levels

## RAID 1+0 (RAID 10)

- Combina *mirror* y *stripe*
- También se habla de RAID 0+1
- Requiere al menos 4 discos
- Soporta un fallo doble según qué discos fallen



# RAID levels (infrecuentes)

## RAID 2

- Emplea códigos de corrección de errores (Hamming)
- Reparte el fichero por los discos (mejora velocidad)
- La rotación de los discos está sincronizada
- Soporta el fallo de un disco
- No se emplea (códigos de corrección ya emplean los discos)

## RAID 3

- Los datos de un bloque están distribuidos a nivel de byte entre los discos
- Una operación I/O emplea todos los discos simultáneamente
- Mejora la velocidad para aplicaciones con un solo flujo de lectura o escritura
- Uso infrecuente (obsoleto)

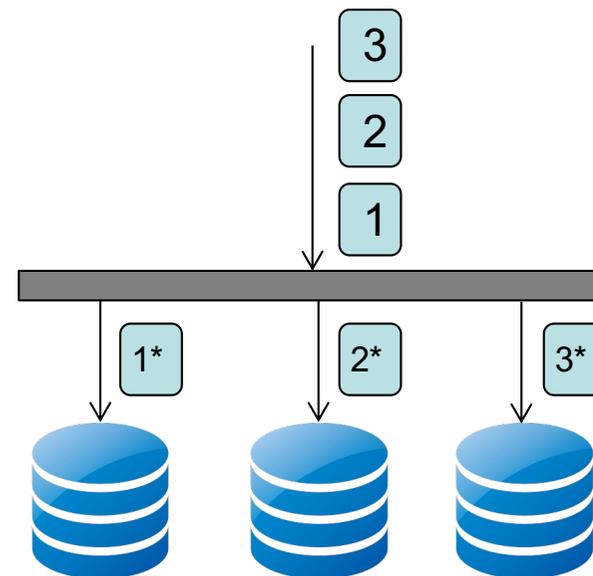
## RAID 4

- Se hace *striping* a nivel de bloques
- Se emplea un disco para paridad
- Mejora la velocidad de lectura porque los datos están repartidos
- En escritura sufre bloqueo pues al tener que escribir la paridad se tiene que hacer en serie para las diferentes peticiones
- Uso infrecuente

# RAID levels

## RAID 5

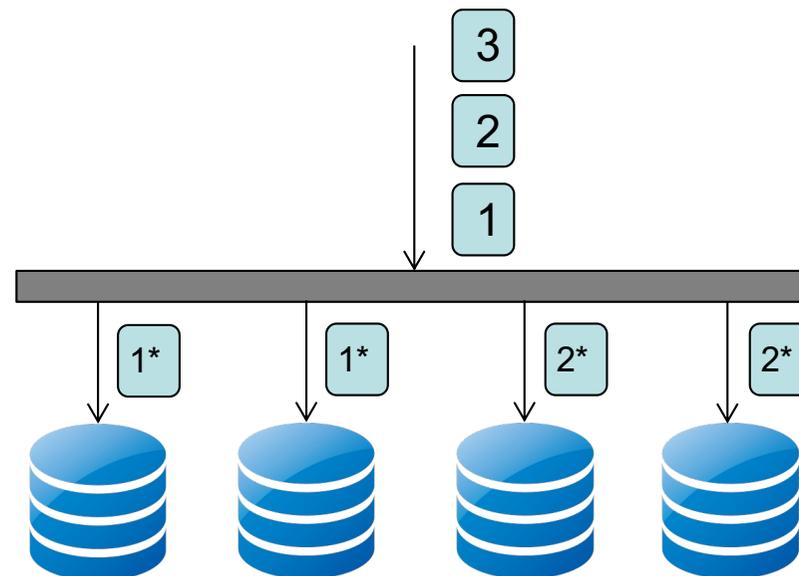
- Se hace *striping* a nivel de bloques
- Se guarda paridad pero no está en el mismo disco la paridad de todos los bloques sino que se reparte por los discos
- Mejora la velocidad porque los datos y la paridad están repartidos
- Menor probabilidad de coincidir múltiples operaciones en el mismo disco cuantos más discos (y así mayor velocidad)



# RAID levels

## RAID 6

- Requiere al menos 4 discos
- Se calcula doble paridad, distribuida por los discos
- Sobrevive a fallos dobles



upna

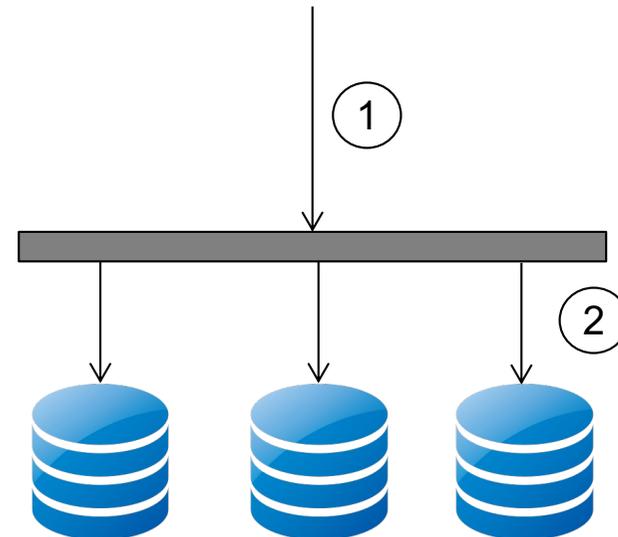
Universidad Pública de Navarra  
Nafarroako Unibertsitate Publikoa

**Redes de Nueva Generación**  
*Área de Ingeniería Telemática*

# RAID

# Interfaces

- Tenemos 2 interfaces
- El primero de ellos (*front-end*) es desde el host (el ordenador) a la controladora
- El segundo (*back-end*) es desde ésta a los discos
- Cuando el disco es interno, simplemente no existe el *front-end*
- El acceso desde el host puede ser a bloques, a ficheros o a registros



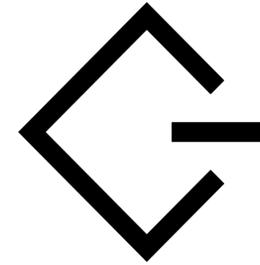
upna

Universidad Pública de Navarra  
Nafarroako Unibertsitate Publikoa

**Redes de Nueva Generación**  
*Área de Ingeniería Telemática*

# Acceso a bloques

# SCSI

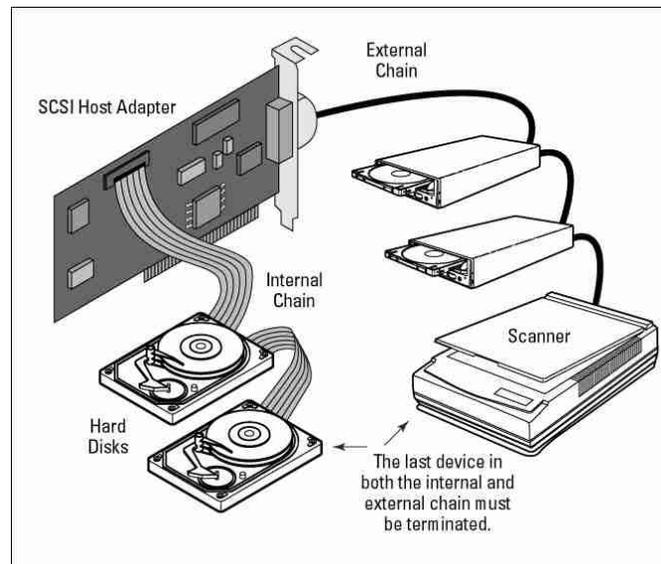


- *Small Computer System Interface*
- Desarrollado por el *International Committee for Information Technology Standards (INCITS)* en los 80s
- Define cómo transferir datos entre ordenadores y periféricos
- Se transfiere a nivel de bloque
- Eso implica comandos, protocolos e interfaces físicos
- Los periféricos más habituales son discos duros pero también ha habido impresoras, scanners, etc.



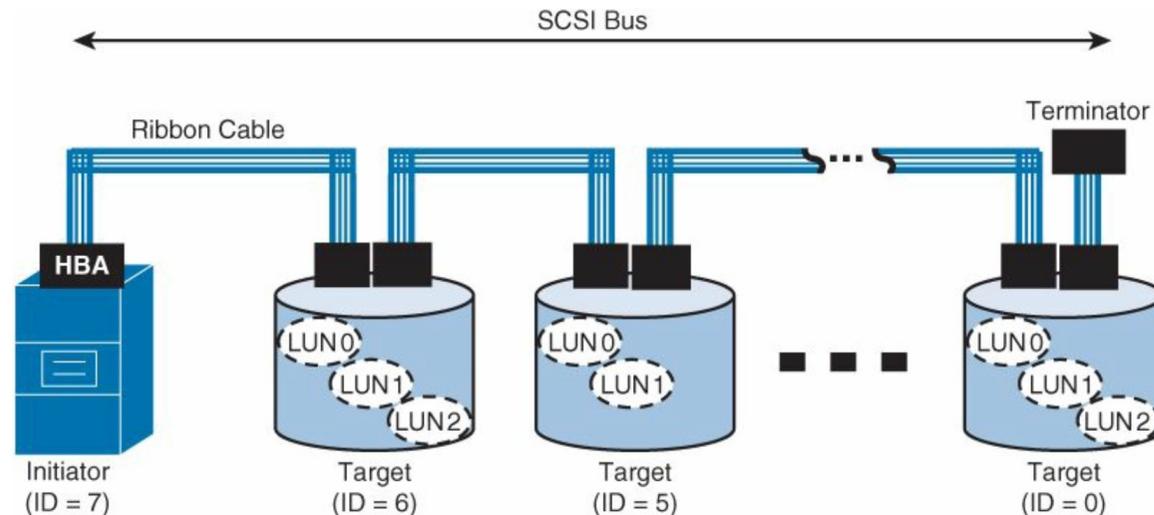
# SCSI bus

- El medio físico es un bus paralelo
- La comunicación es half-duplex
- Dispositivos encadenados
- El bus requiere un terminador
- Un elemento es el “*Initiator*”, normalmente la controladora en el ordenador que accede a los periféricos
- Cada dispositivo (*target*) tiene un identificador numérico que implica también la prioridad del mismo



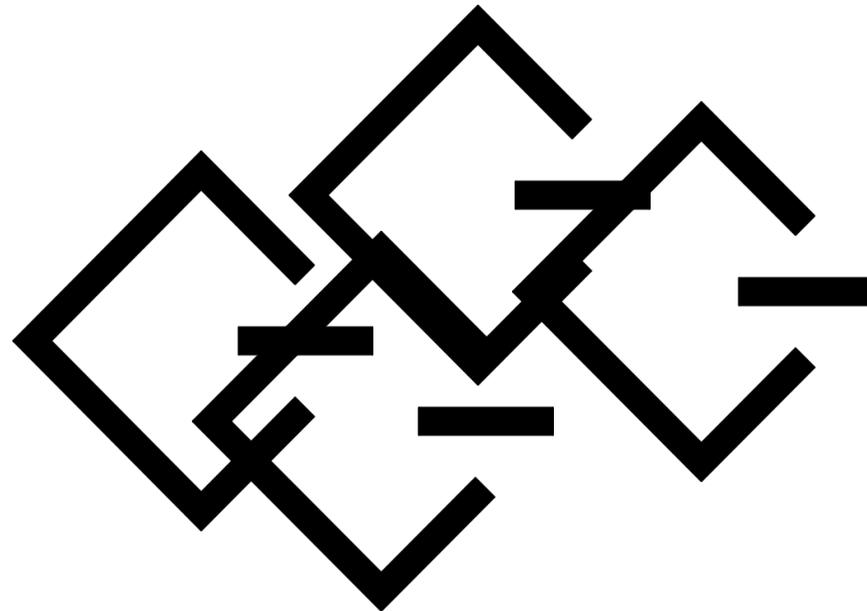
# SCSI bus

- El iniciador direcciona unidades lógicas (“*logical units*”)
- Cada una de las cuales se identifica con un *Logical Unit Number (LUN)*
- Discos duros pueden tener más de un LUN
- La controladora SCSI es lo que se llama un *Host Bus Adapter (HBA)*
- Los comandos principales en el bus son simplemente “Read” y “Write” aunque hay otros para diagnóstico, formateo, etc.



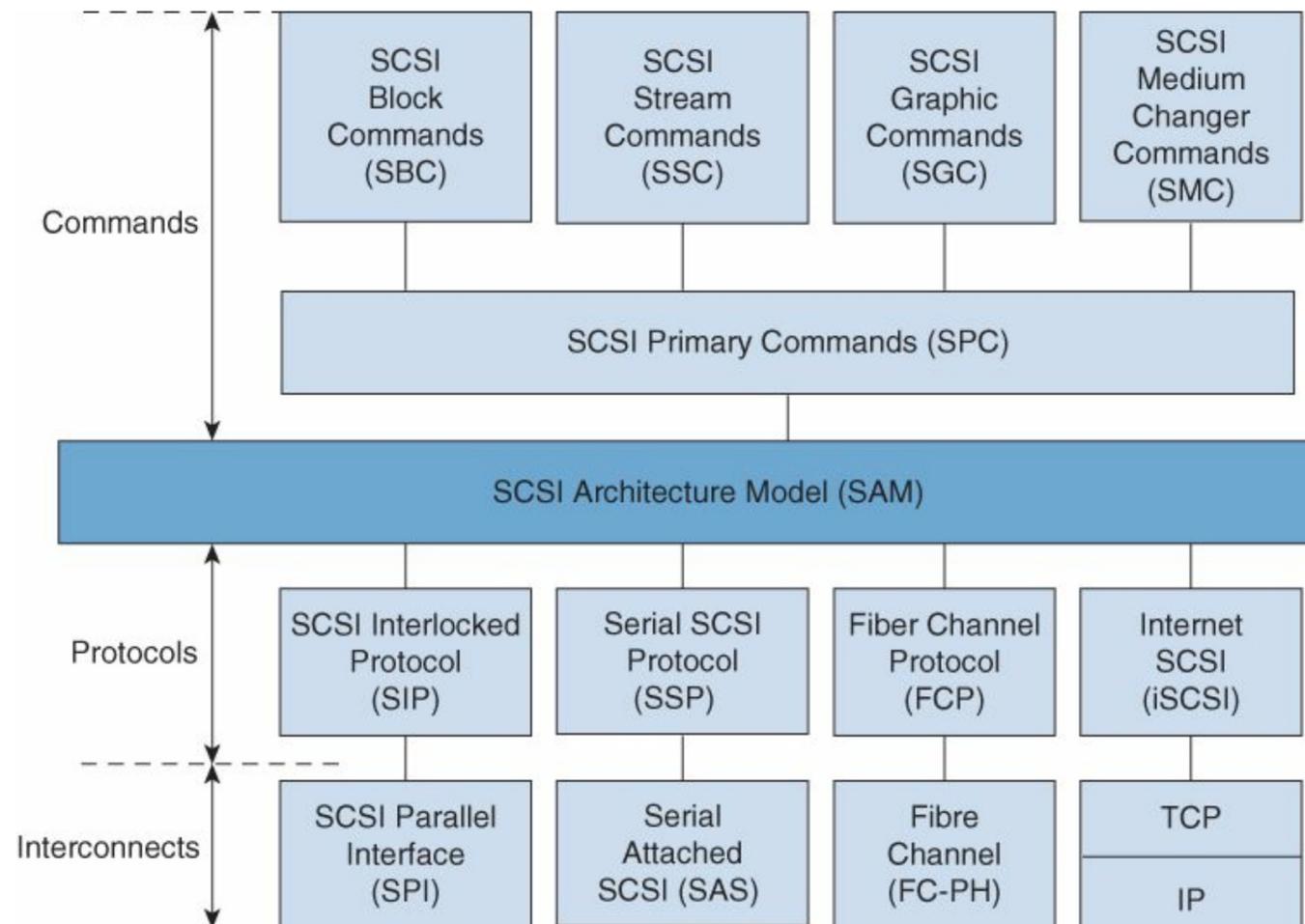
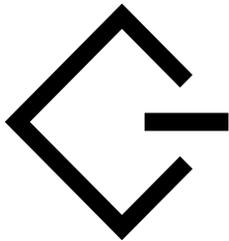
# Versiones de SCSI

- SCSI-1 o el SCSI original
  - 5 Mbytes/s
  - Hasta 8 dispositivos en el bus (numerados de 0 a 7) incluyendo el iniciador
  - Conector Centronics, 50 pines
- Versiones posteriores: SCSI-2, SCSI-3, Fast SCSI, Wide SCSI, Ultra SCSI, Ultra Wide SCSI, Ultra 160 SCSI, Ultra 320 SCSI...
- Aumentan la anchura del bus, la velocidad, el número de dispositivos en el mismo
- Versiones hasta 640Mbps, 16 dispositivos, 25 metros
- Se encuentra con limitaciones debidas al cable paralelo



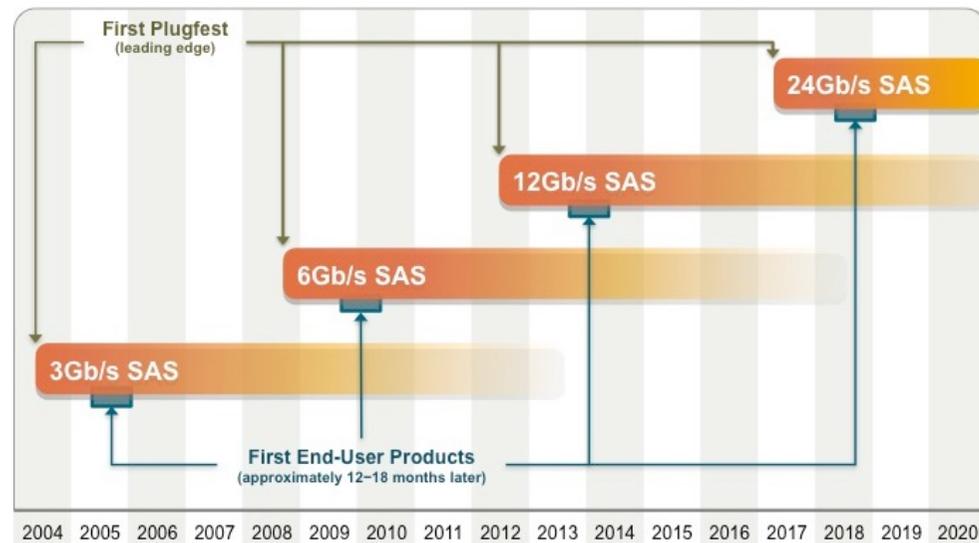
# SCSI hoy en día

- Se ha independizado el protocolo (los comandos) del interfaz físico
- Ha evolucionado hacia nuevos medios físicos
- (...)



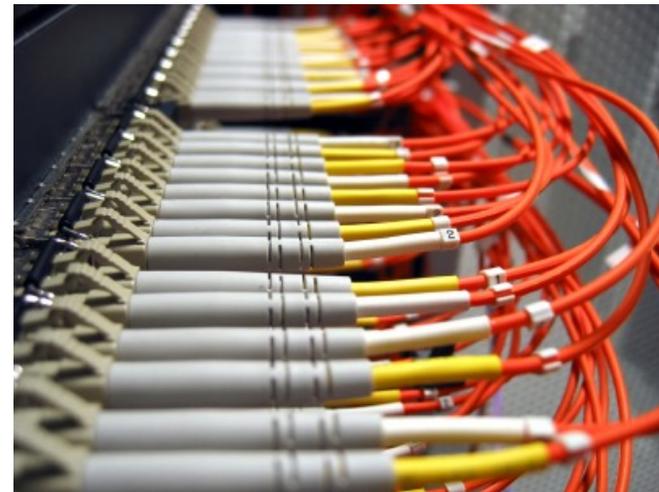
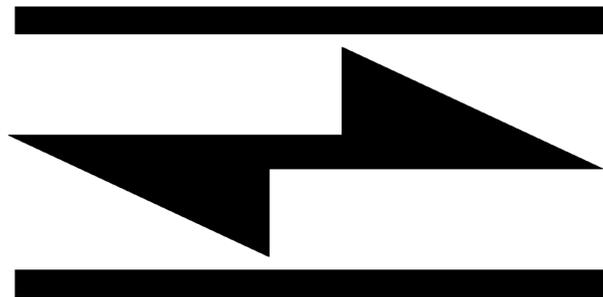
# SCSI hoy en día

- Se ha independizado el protocolo (los comandos) del interfaz físico
- Ha evolucionado hacia nuevos medios físicos
- SAS = *Serial Attached SCSI*
  - Se abandona el cable paralelo por un medio serie
  - Se abandona el bus por enlace punto-a-punto
  - Hoy en día tasas de hasta 12 Gbps y hasta 10 m
- (...)



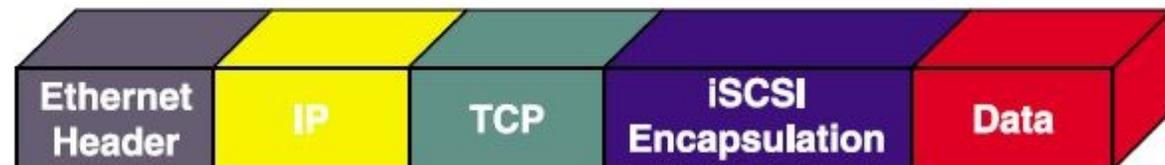
# SCSI hoy en día

- Se ha independizado el protocolo (los comandos) del interfaz físico
- Ha evolucionado hacia nuevos medios físicos
- *SAS = Serial Attached SCSI*
- *Fibre Channel*
  - Normalmente sobre fibra (no necesariamente)
  - Una tecnología de red principalmente para almacenamiento
  - Que transporta comandos SCSI
  - Velocidades hoy en día de decenas de Gbps, distancias de kms
  - Transportable sobre WAN
- (...)



# SCSI hoy en día

- Se ha independizado el protocolo (los comandos) del interfaz físico
- Ha evolucionado hacia nuevos medios físicos
- *SAS = Serial Attached SCSI*
- *Fibre Channel*
- *iSCSI (Internet SCSI)*
  - Comandos SCSI sobre una conexión TCP



upna

Universidad Pública de Navarra  
Nafarroako Unibertsitate Publikoa

**Redes de Nueva Generación**  
*Área de Ingeniería Telemática*

# Acceso a bloques

upna

Universidad Pública de Navarra  
Nafarroako Unibertsitate Publikoa

**Redes de Nueva Generación**  
*Área de Ingeniería Telemática*



# SSD



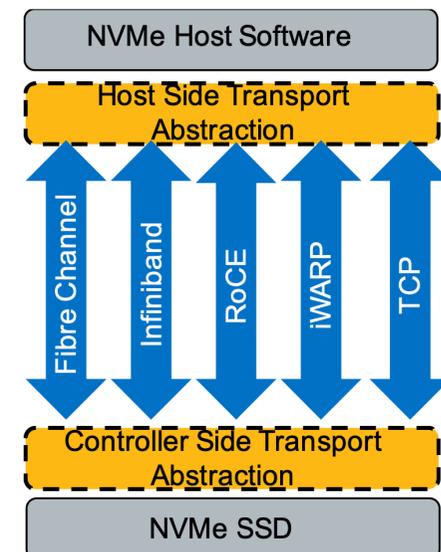
# Discos SSD

- *Solid State Disk*
- En realidad no es un “disco” sino memoria, pero la utilizamos como almacenamiento permanente
- NAND Flash
- No tiene sentido hablar de tiempo de posicionar la cabeza lectora o de esperar a que rote el disco
- Estamos hablando de tiempos de acceso en el orden de los microsegundos
- Hasta hace unos pocos años no han tenido un precio que fuera compatible con sus ventajas
- Envejecimiento y limitaciones de escritura/borrado

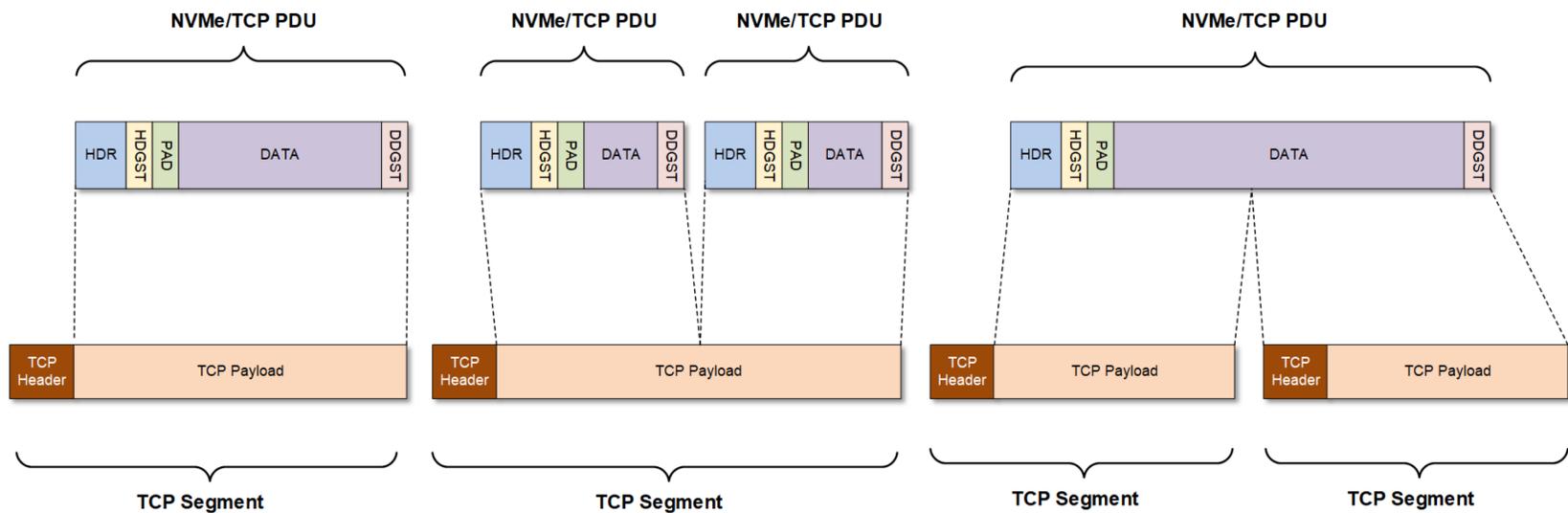
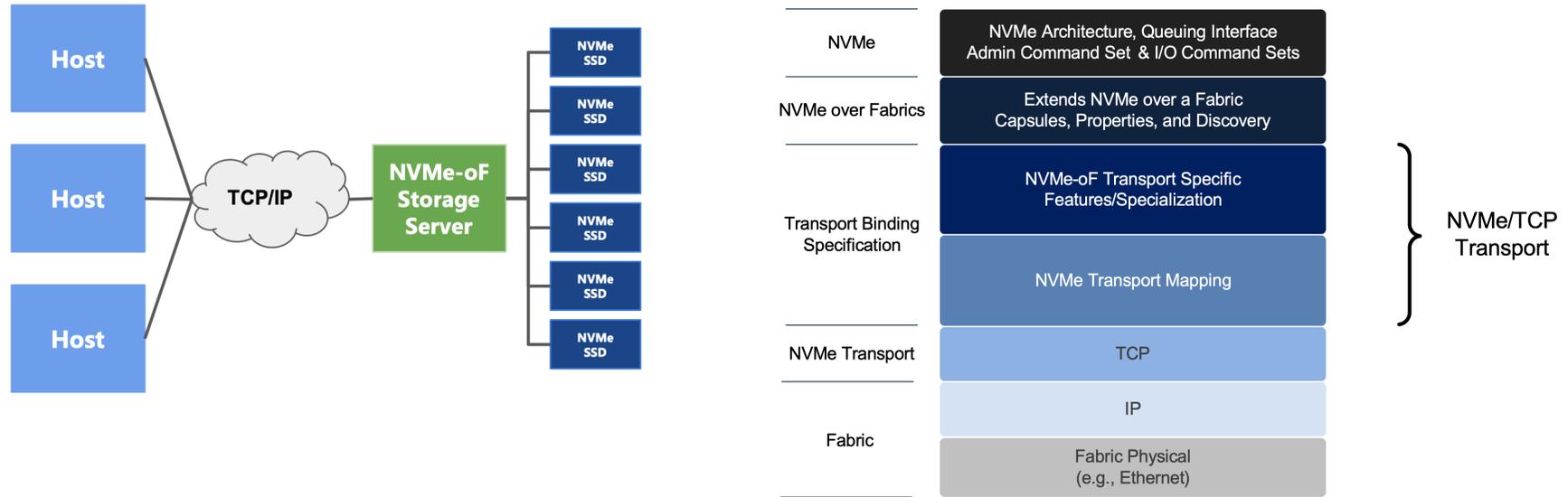


# NVMe

- Non-Volatile Memory express
- Protocolos de transporte y acceso a almacenamiento para SSD
- Conexión directa a CPU a través de PCIe
- Aumenta el paralelismo en las peticiones al disco (64K colas de 64K peticiones cada una)
- Menor latencia
- NVMe-oF : NVMe over Fabrics
  - Uso de DMA (zero-copy)
  - NVMe sobre RDMA (Infiniband, RoCE, iWARP)
  - NVMe/FC : comandos NVMe sobre FC en lugar de SCSI
  - NVMe/TCP



# NVMe/TCP



upna

Universidad Pública de Navarra  
Nafarroako Unibertsitate Publikoa

**Redes de Nueva Generación**  
*Área de Ingeniería Telemática*



# SSD



upna

Universidad Pública de Navarra  
Nafarroako Unibertsitate Publikoa

**Redes de Nueva Generación**  
*Área de Ingeniería Telemática*

# Discos

# Disco SATA

Especificaciones	8 TB	6 TB	4 TB	3 TB	2TB
Número de modelo estándar	ST8000DM004	ST6000DM003	ST4000DM004	ST3000DM007	ST2000DM005
Bytes por sector	4.096	4096	4.096	4 096	4 096
<b>Rendimiento</b>					
Interfaz	SATA a 6 Gb/s	SATA a 6 Gb/s	SATA a 6 Gb/s	SATA a 6 Gb/s	SATA a 6 Gb/s
Compatible con transferencias de datos SATA (Gb/s)	6,0 / 3,0 / 1,5	6,0/3,0/1,5	6,0/3,0/1,5	6,0 / 3,0 / 1,5	6,0/3,0/1,5
Velocidad de transferencia de datos sostenida máxima, lectura OD (MB/s)	190MB/s	185MB/s	190MB/s	185MB/s	190MB/s
Caché multisegmentada (MB)	256	256	256	256	256
Velocidad de giro (rpm)	5.400	5.400	5.400	5.400	5.400
Tecnología de grabación	SMR	SMR	SMR	SMR	SMR
<b>Fiabilidad/integridad de datos</b>					
Ciclos de carga/descarga	600.000	600.000	600.000	600.000	600.000
Errores de lectura no recuperables por bits leídos, máx.	1 por 10E14	1 por 10E14	1 por 10E14	1 por 10E14	1 por 10E14
Horas de funcionamiento (al año)	2.400	2.400	2.400	2.400	2.400
Límite del índice de carga de trabajo (TB/año)	55	55	55	55	55
Garantía, limitada (años)	2	2	2	2	2



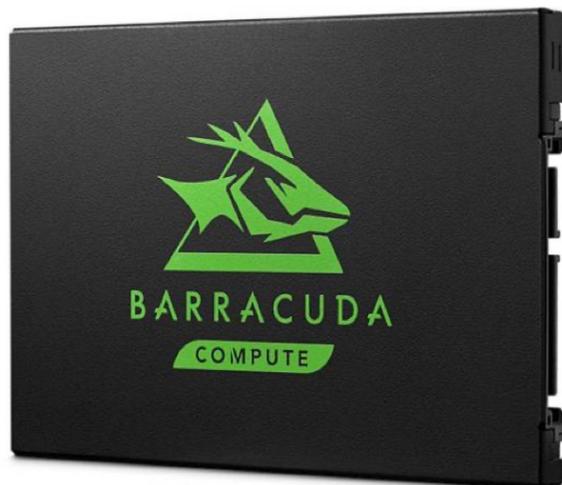
# Disco SATA (Pro)

Especificaciones	14 TB	12 TB	12 TB	10 TB	10 TB
Números de modelo estándar	ST14000DM001	ST12000DM001	ST12000DM0007	ST10000DM001	ST10000DM0004
Bytes por sector	512	512	512	512	512
<b>Rendimiento</b>					
Interfaz	SATA a 6 Gb/s	SATA a 6 Gb/s	SATA a 6 Gb/s	SATA a 6 Gb/s	SATA a 6 Gb/s
Compatible con transferencias de datos SATA (Gb/s)	6,0/3,0/1,5	6,0/3,0/1,5	6,0/3,0/1,5	6,0/3,0/1,5	6,0/3,0/1,5
Velocidad de transferencia de datos sostenida máxima, lectura OD (MB/s)	250MB/s	250MB/s	250MB/s	220MB/s	220MB/s
Caché multisegmentada (MB)	256	256	256	256	256
Tecnología de grabación	CMR	CMR	CMR	CMR	CMR
<b>Fiabilidad/integridad de datos</b>					
Ciclos de carga/descarga	300.000	300.000	300.000	300.000	300.000
Errores de lectura no recuperables por bits leídos, máx.	1 por cada 10E15	1 por cada 10E15	1 por cada 10E15	1 por 10E15	1 por cada 10E15
Horas de funcionamiento (al año)	8.760	8.760	8.760	8.760	8.760
Servicios de recuperación de datos y Rescue	Sí	Sí	Sí	Sí	Sí
Límite de carga de trabajo (TB/año) <sup>1</sup>	300	300	300	300	300
Garantía limitada (años)	5	5	5	5	5



# Disco SSD SATA

Especificaciones	2 TB	1 TB	500 GB	250 GB
Número de modelo	ZA2000CM10003	ZA1000CM10003	ZA500CM10003	ZA250CM10003
Interfaz	SATA a 6 Gb/s			
Memoria NAND Flash	3D TLC	3D TLC	3D TLC	3D TLC
Tamaño	2,5 in x 7 mm			
<b>Rendimiento</b>				
Lectura secuencial (máxima, MB/s), 128 KB <sup>1</sup>	560	560	560	560
Escritura secuencial (máxima, MB/s), 128 KB <sup>1</sup>	540	540	540	540
Lectura aleatoria (máxima, IOPS), 4 KB QD32 <sup>1</sup>	90 000	90 000	90 000	90 000
Escritura aleatoria (máxima, IOPS), 4 KB QD32 <sup>1</sup>	90 000	90 000	90 000	90 000
<b>Durabilidad/Fiabilidad</b>				
Total de bytes escritos (TB)	1170	600	300	150
Tiempo medio entre fallos (MTBF, horas)	1 800 000	1 800 000	1 800 000	1 800 000
Garantía limitada (años)	5	5	5	5



## 6.3.1 Total Bytes Written (TBW)

TBW (Total Bytes Written) is a measurement of the SSDs' expected lifespan, which represents the amount of data written to the device. To calculate the TBW of a SSD, the following equation is applied:

$$TBW = [(NAND Endurance) \times (SSD Capacity)] / WAF$$

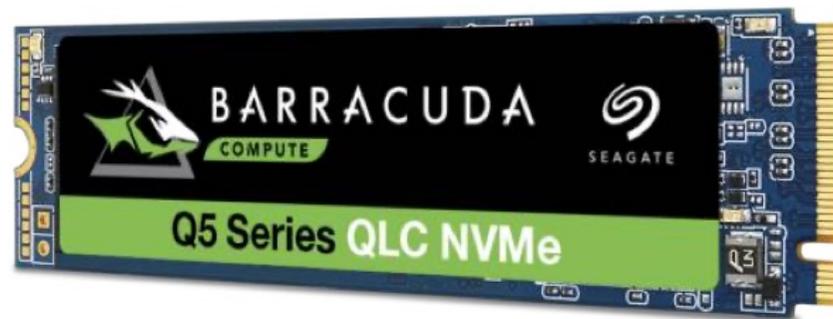
NAND Endurance: NAND endurance refers to the P/E (Program/Erase) cycle of a NAND flash. SSD Capacity: The SSD capacity is the specific capacity in total of a SSD.

WAF: Write Amplification Factor (WAF) is a numerical value representing the ratio between the amount of data that a SSD controller needs to write and the amount of data that the host's flash controller writes. A better WAF, which is near 1, guarantees better endurance and lower frequency of data written to flash memory.

TBW in this document is based on JEDEC 219 workload.

# Discos SSD NVMe

Especificaciones	2 TB	1 TB	500 GB
Modelo estándar	ZP2000CV30001	ZP1000CV30001	ZP500CV30001
Interfaz	PCIe <sup>®</sup> Gen3 x4 NVMe <sup>™</sup> 1.3	PCIe Gen3 x4 NVMe 1.3	PCIe Gen3 x4 NVMe 1.3
Memoria flash NAND	3D QLC	3D QLC	3D QLC
Tamaño	M.2 2280-S2	M.2 2280-S2	M.2 2280-S2
<b>Rendimiento</b>			
Lectura secuencial (máxima, MB/s), 128 KB <sup>1</sup>	2400	2400	2300
Escritura secuencial (máxima, MB/s), 128 KB <sup>1</sup>	1800	1700	900
<b>Durabilidad/fiabilidad</b>			
Total de bytes escritos (TB)	531	274	119
Tiempo medio entre fallos (MTBF, horas)	1 800 000	1 800 000	1 800 000
Garantía limitada (años)	3	3	3
Servicios de recuperación de datos Rescue para recuperación de datos (años)	1	1	1



upna

Universidad Pública de Navarra  
Nafarroako Unibertsitate Publikoa

**Redes de Nueva Generación**  
*Área de Ingeniería Telemática*

# Discos