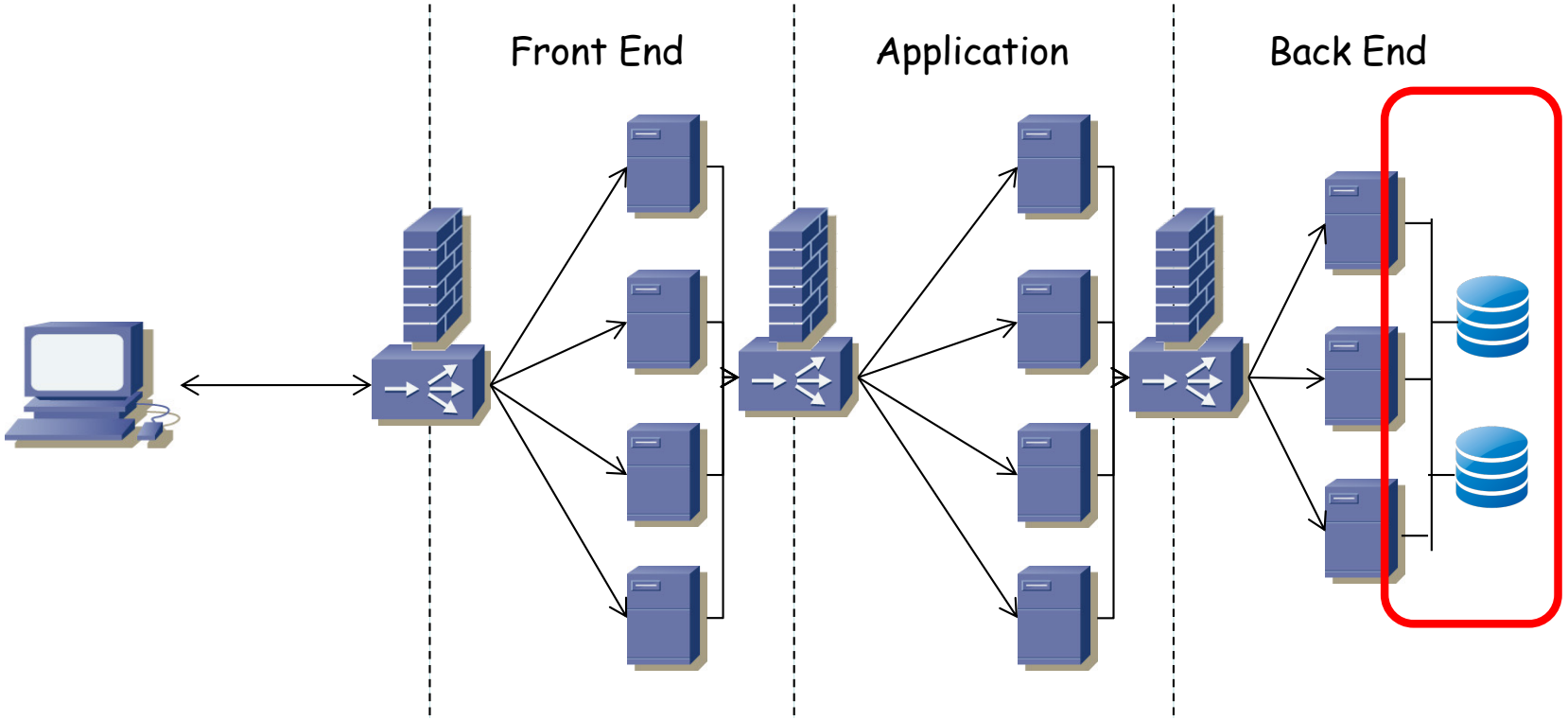
upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

Almacenamiento

Servicios y almacenamiento

- Por supuesto todos los servidores lo necesitan
- Lo tienen local (puede que lo usen o no)
- Pero hemos visto que principalmente está en la capa más profunda del *backend*
- Veremos ahora cómo se implementa

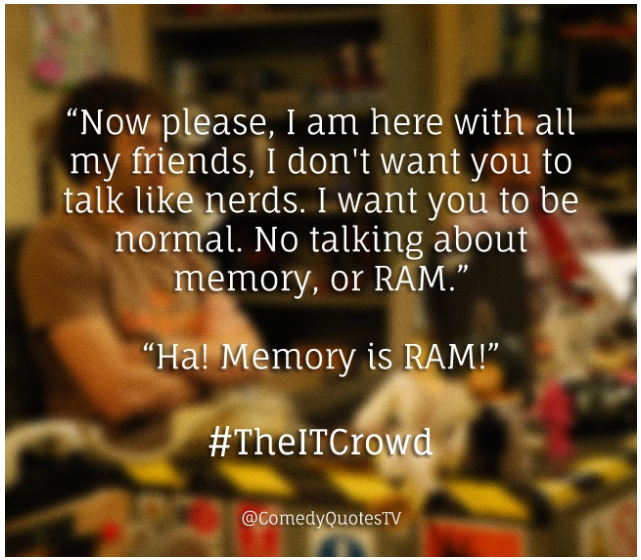


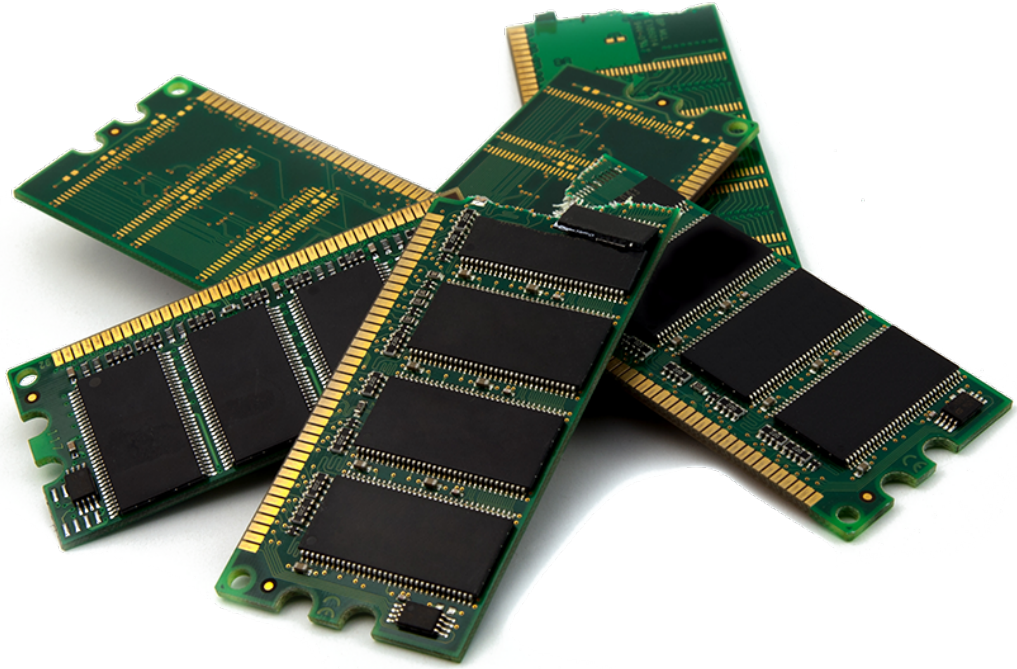
upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

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

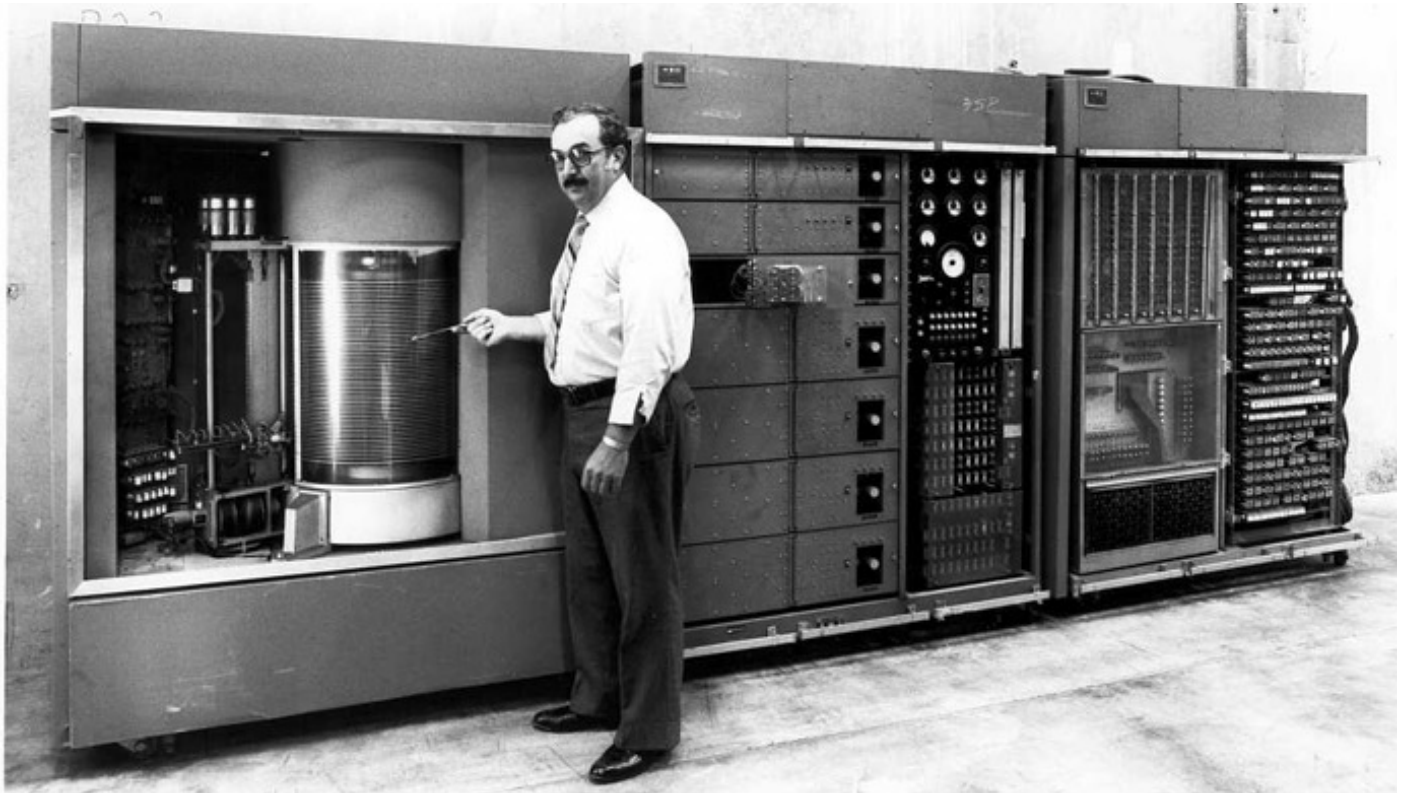
El disco duro

Almacenamiento secundario

- Mainframes empleaban cintas magnéticas

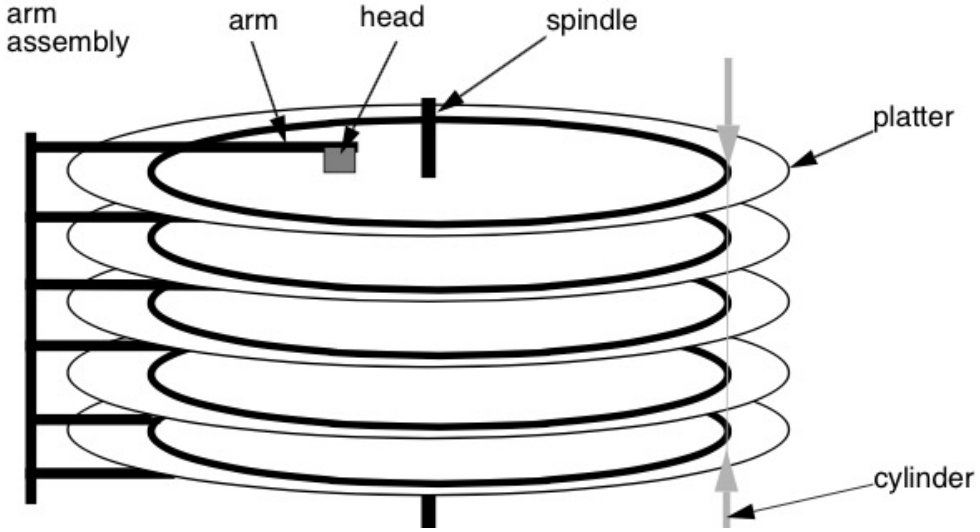
Almacenamiento secundario

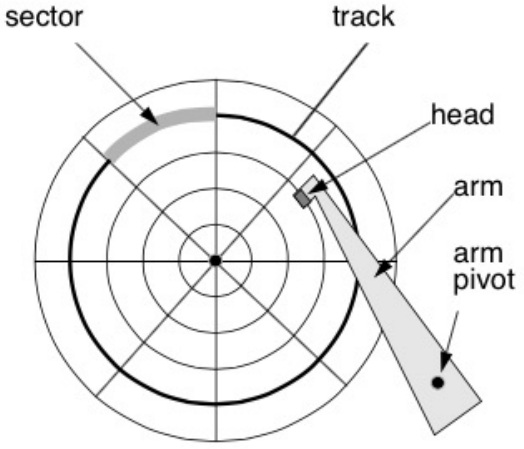
- A mediados de los 50 surge el disco duro o *“hard disk drive”*
- Sigue siendo magnético pero en lugar de una cinta son platos



Arquitectura del disco

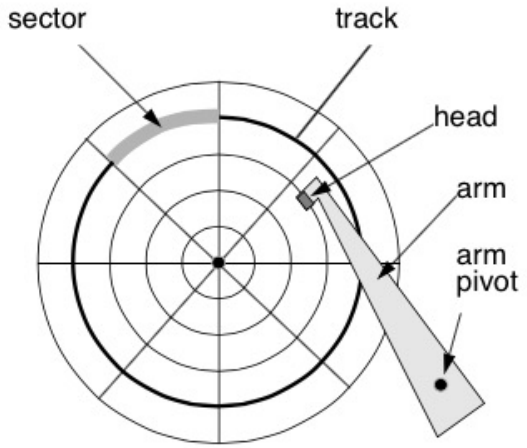
- Platos (*platters*): material magnético
- Pistas (*tracks*): trayectoria circular
- Sectores: unidad mínima direccionable, todos del mismo tamaño, tradicionalmente 512 Bytes (hay ahora unidades con 4 KiB)
- Cilindros: pila vertical de pistas
- El tamaño del disco (en pulgadas) condiciona su capacidad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





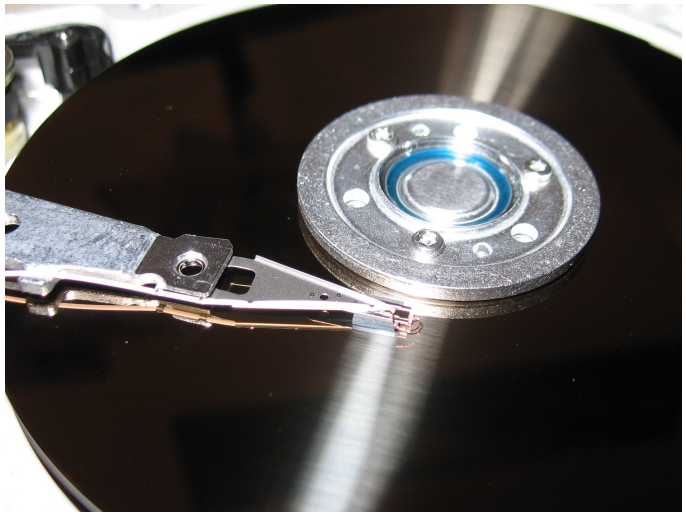
Rotación del disco

- 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

RAID

RAID

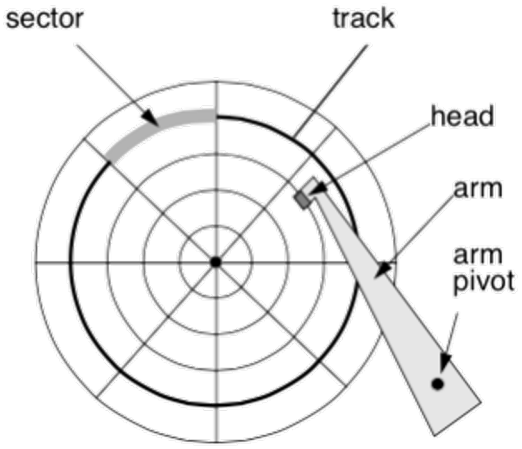
- *Redundant Array of Independent (Inexpensive) Disks*
- Varios discos que de cara al usuario (el servidor) se comportan como un solo volumen
- Los diferentes tipos de RAID se denominan mediante un “nivel”
- RAID level 0, RAID level 1, etc
- En comparación con *Just a Bunch Of Disks* (JBOD)

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 irrecuperable
- Cualquier número de discos (mayor que 1)

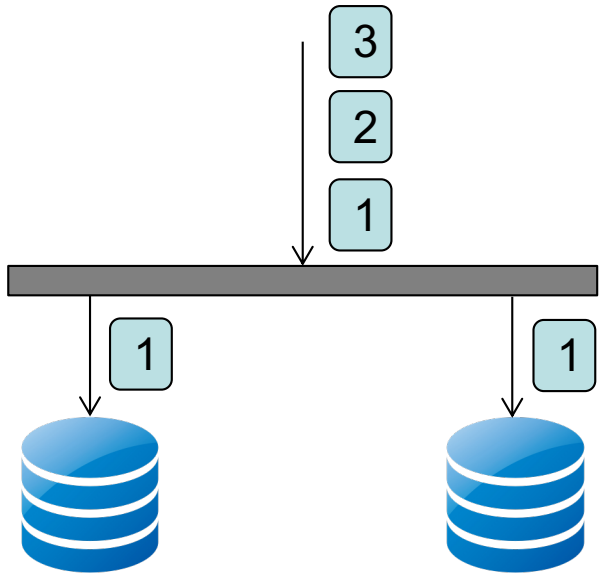
- *Slicing*
  - La velocidad lineal es superior en las pistas exteriores
  - Se puede aumentar la velocidad creando el RAID empleando solo esos sectores



RAID levels

RAID 1

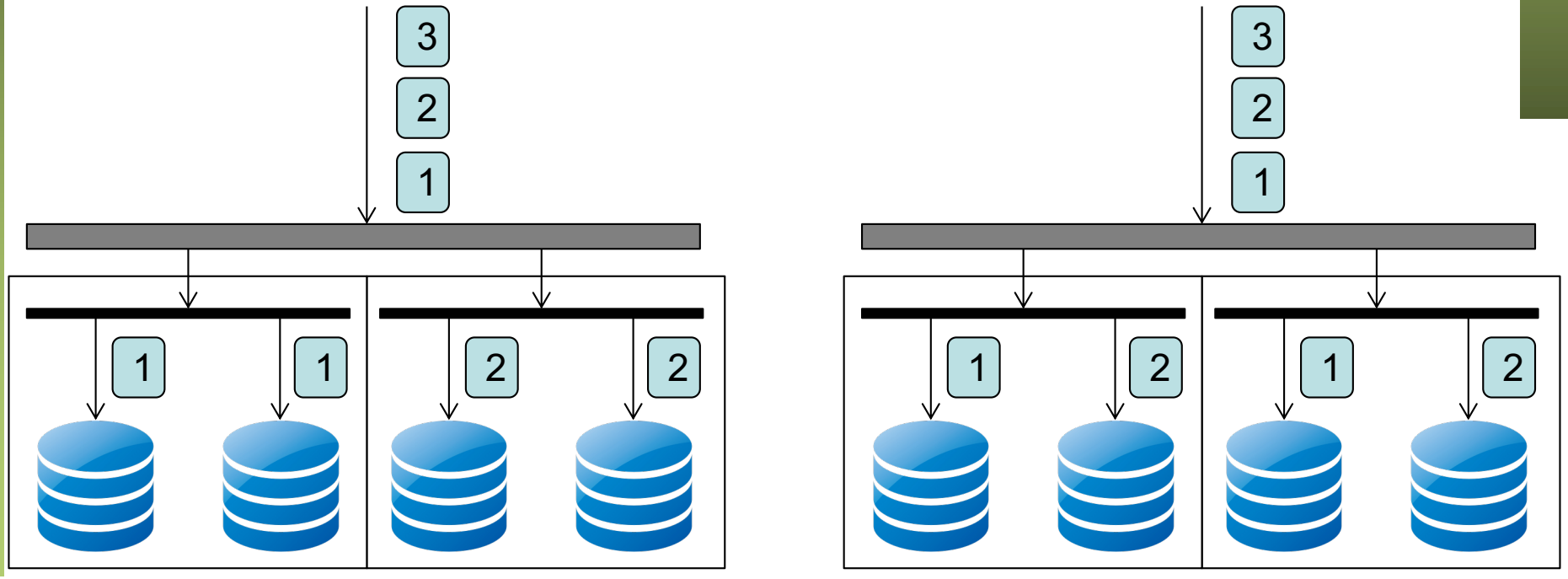
- *Mirroring*
- Los datos se replican
- Requiere al menos 2 discos
- No mejora la velocidad pero sí da protección
- Ante un fallo en un disco el RAID puede seguir funcionando
- Se puede sustituir el disco defectuoso y la controladora reconstruye el *mirror*
- La reconstrucción reduce el rendimiento del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

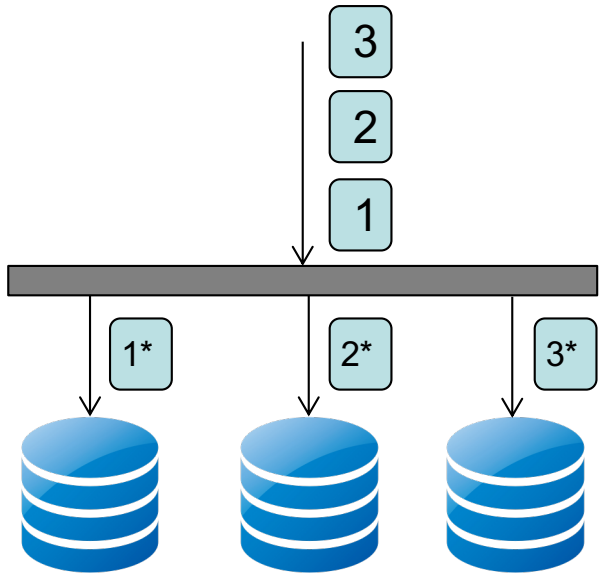
RAID 4

- 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

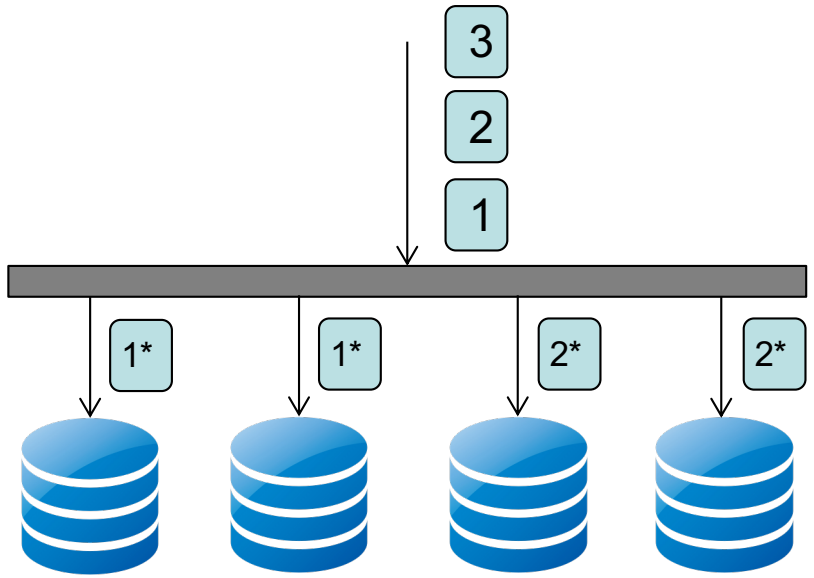
- Se hace *striping* a nivel de bloques
- Se guarda paridad pero no está en el mismo disco la paridad de todos los bloques sino que se reparte por los discos
- Mejora la velocidad porque los datos y la paridad están repartidos
- Menor probabilidad de coincidir múltiples operaciones en el mismo disco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



RAID levels

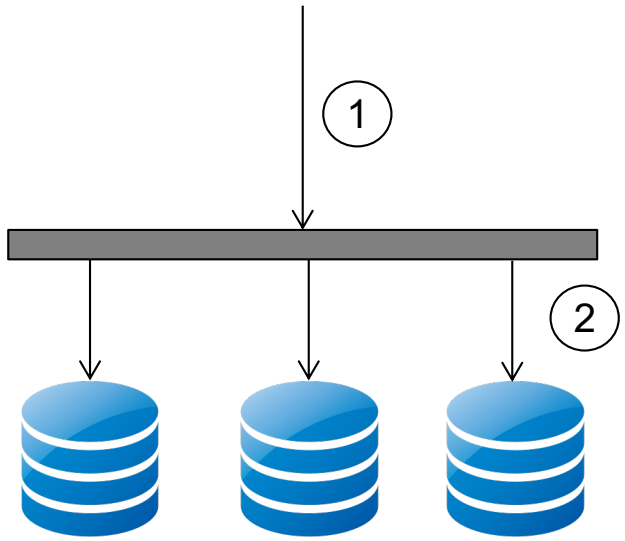
RAID 6

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



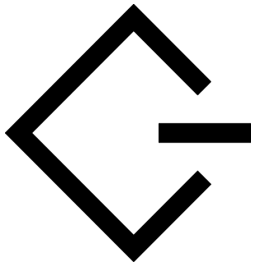
Interfaces

- 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

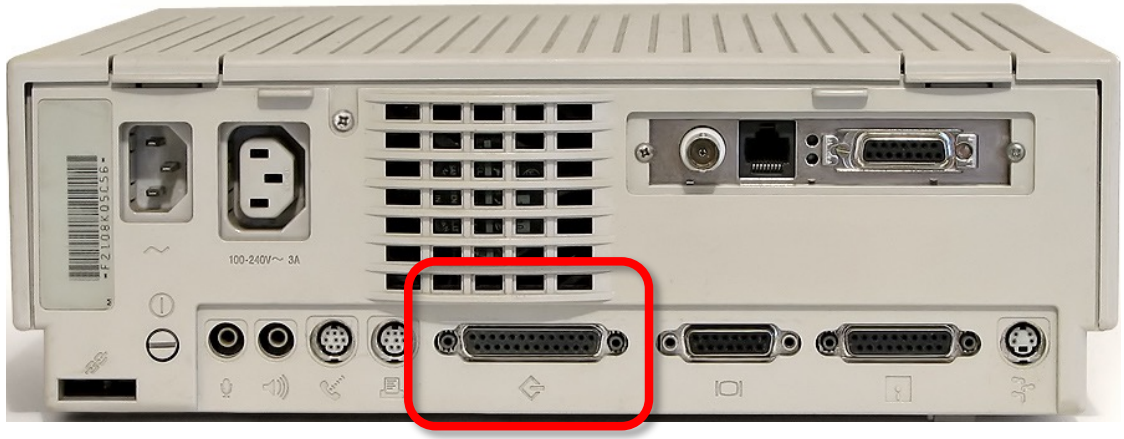


Acceso a bloques

SCSI

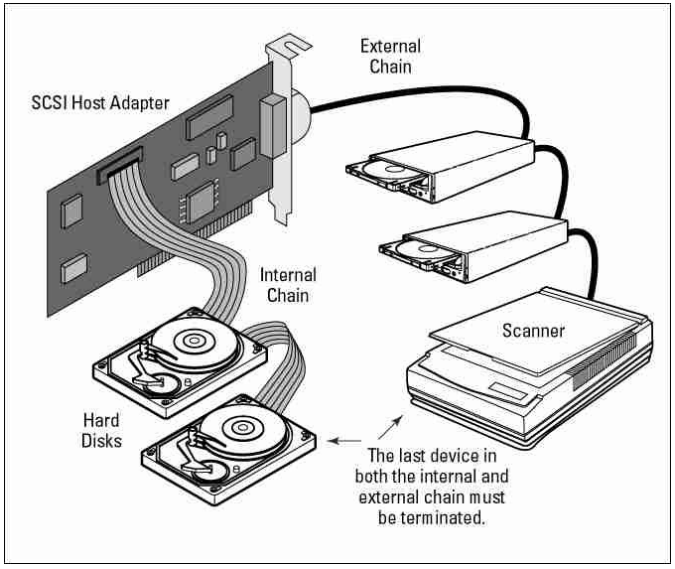


- *Small Computer System Interface*
- Desarrollado por el *International Committee for Information Technology Standards (INCITS)* en los 80s
- Define cómo transferir datos entre ordenadores y periféricos
- Se transfiere a nivel de bloque
- Eso implica comandos, protocolos e interfaces físicos
- Los periféricos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



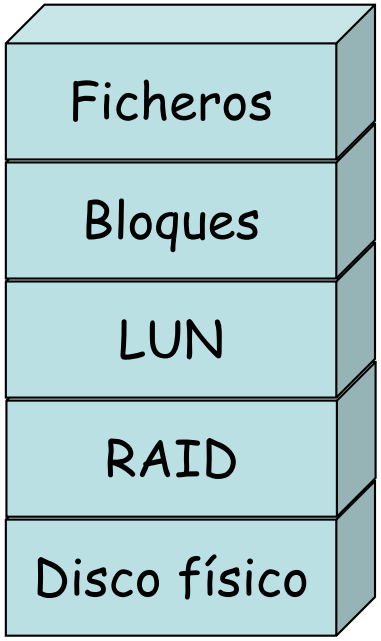
SCSI bus

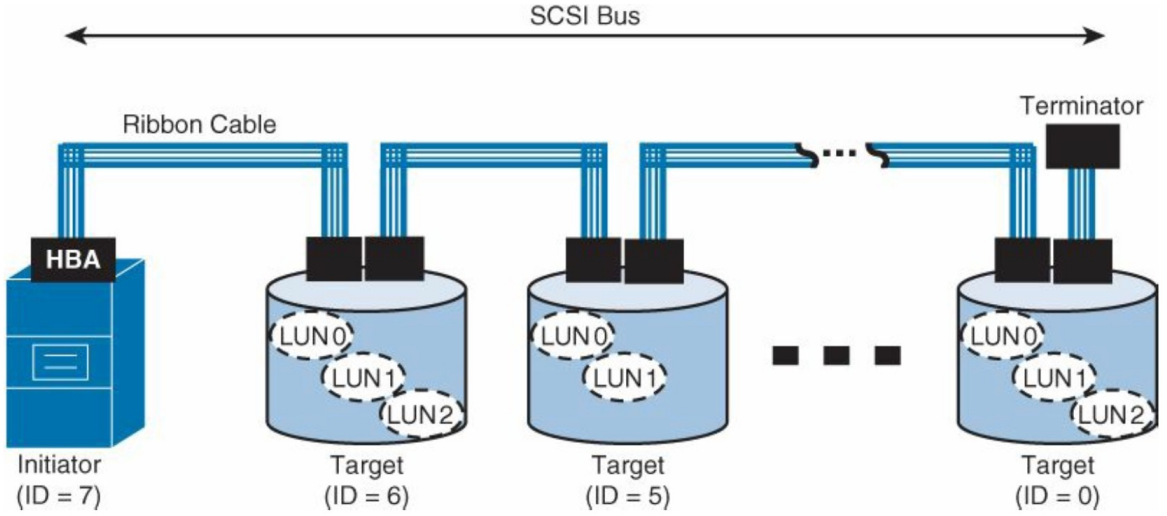
- El medio físico es un bus paralelo
- La comunicación es half-duplex
- Dispositivos encadenados
- El bus requiere un terminador
- Un elemento es el "*Initiator*", normalmente la controladora en el ordenador que accede a los periféricos
- Cada dispositivo (*target*) tiene un identificador numérico que 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 Adapteer (HBA)*
- Los comandos principales en el bus son simplemente “Read” y “Write” aunque hay otros para diagnóstico, formateo, etc.

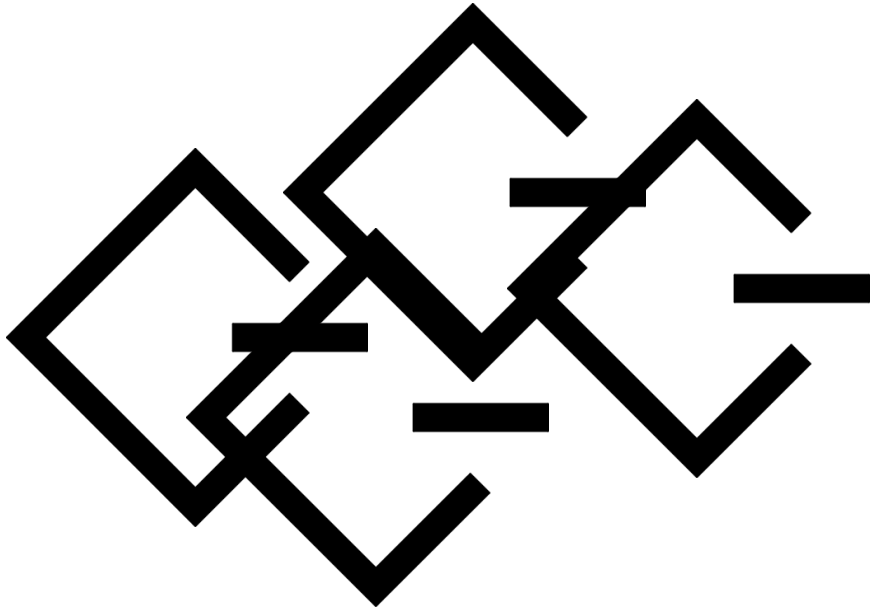




G.A.A.Santana, “Data Center Virtualization Fundamentals”, Cisco Press

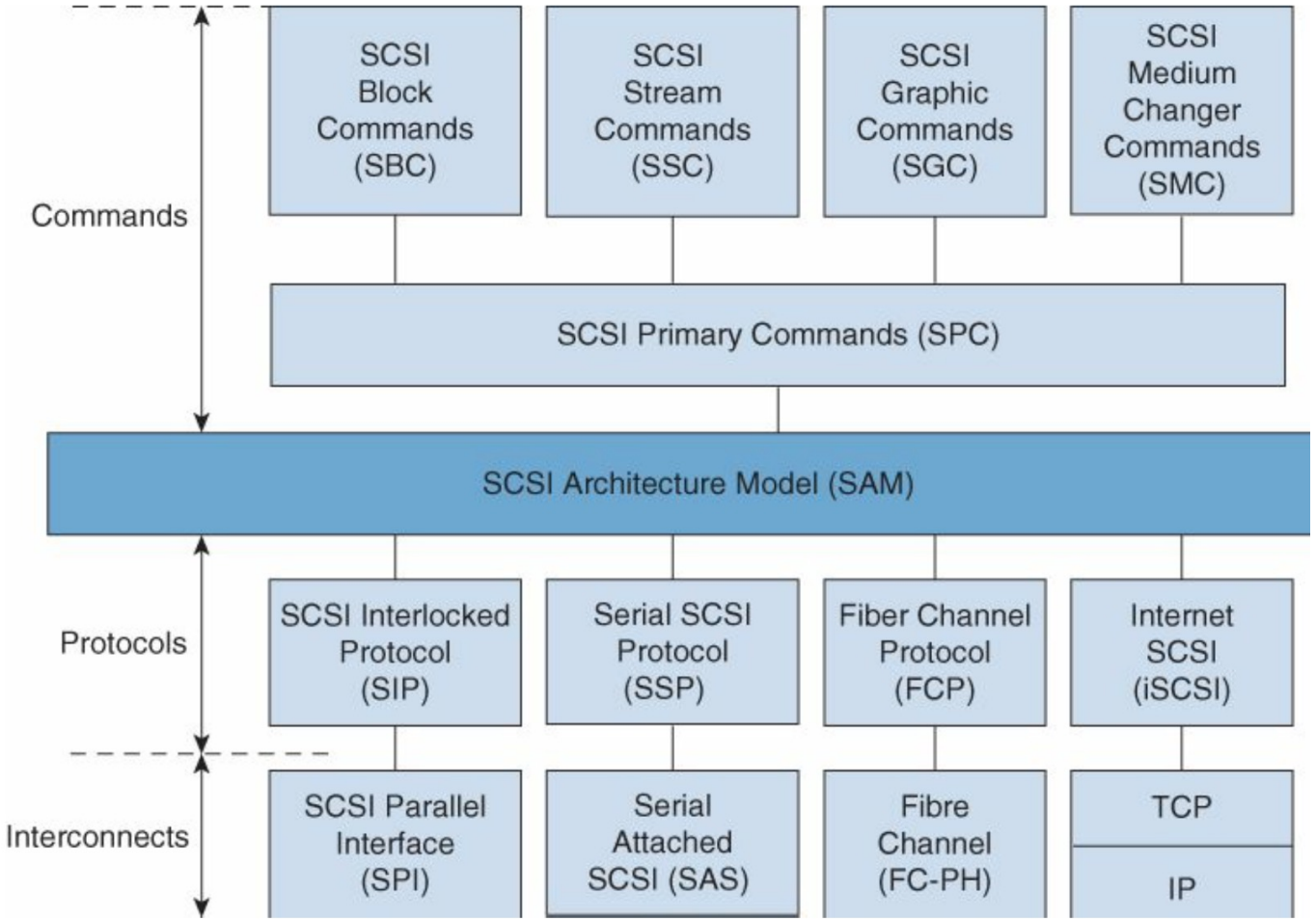
Versiones de SCSI

- 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

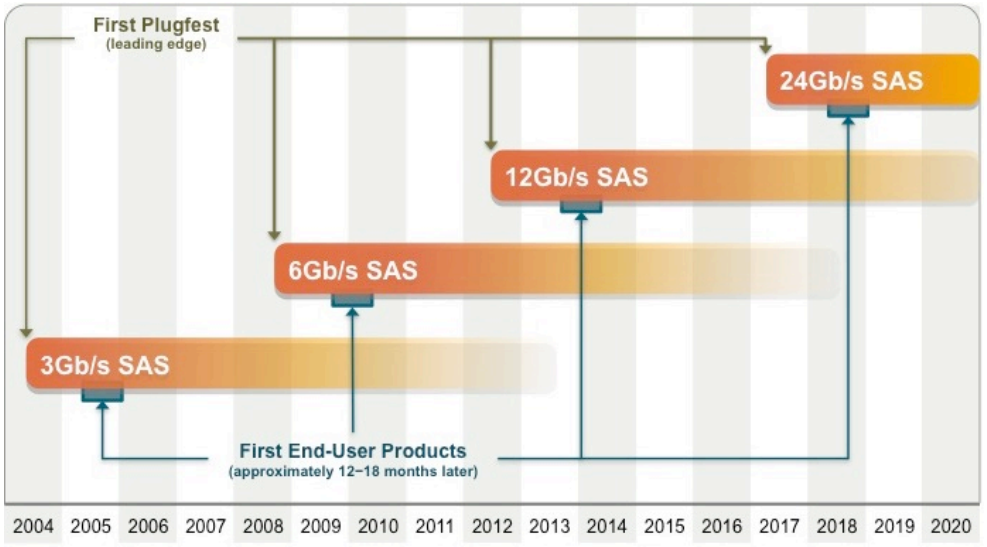
- Se ha independizado el protocolo (los comandos) del interfaz físico
- Ha evolucionado hacia nuevos medios físicos
- (…)



G.A.A.Santana, “Data Center Virtualization Fundamentals”, Cisco Press

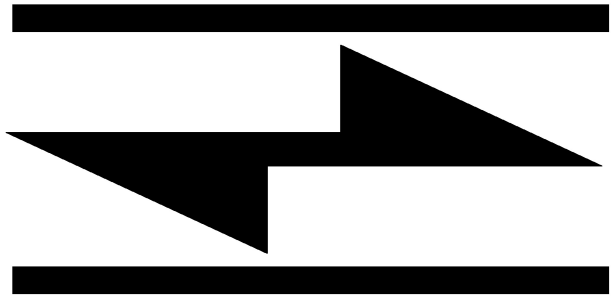
SCSI hoy en día

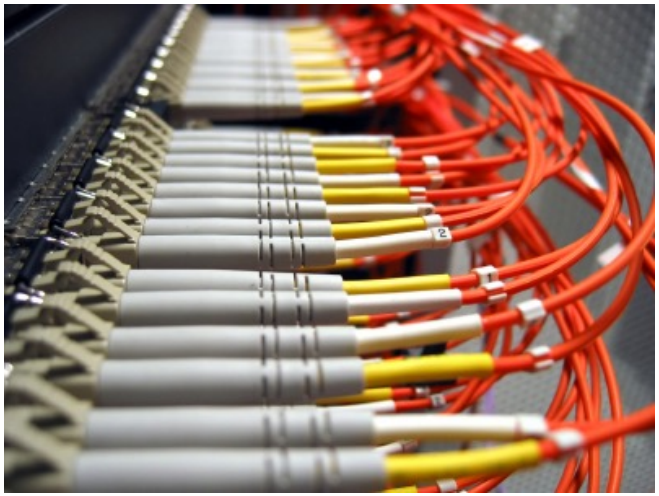
- Se ha independizado el protocolo (los comandos) del interfaz físico
- Ha evolucionado hacia nuevos medios físicos
- SAS = *Serial Attached SCSI*
  - Se abandona el cable paralelo por un medio serie
  - Se abandona el bus por enlace punto-a-punto
  - Hoy en día tasas de hasta 12 Gbps y hasta 10 m
-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



SCSI hoy en día

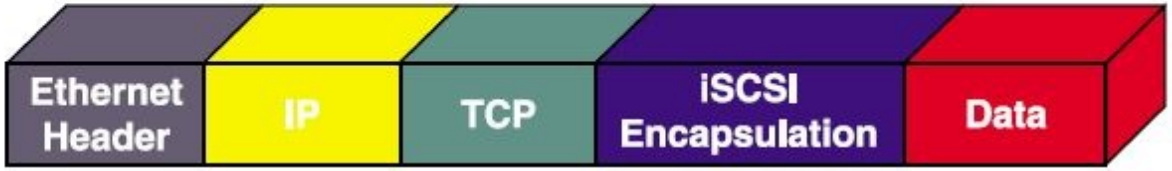
- Se ha independizado el protocolo (los comandos) del interfaz físico
- Ha evolucionado hacia nuevos medios físicos
- SAS = *Serial Attached SCSI*
- *Fibre Channel*
  - Normalmente sobre fibra (no necesariamente)
  - Una tecnología de red principalmente para almacenamiento
  - Que transporta comandos 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

SSD

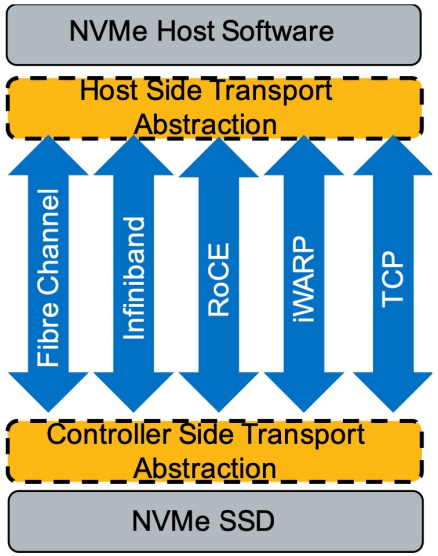
Discos SSD

- *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

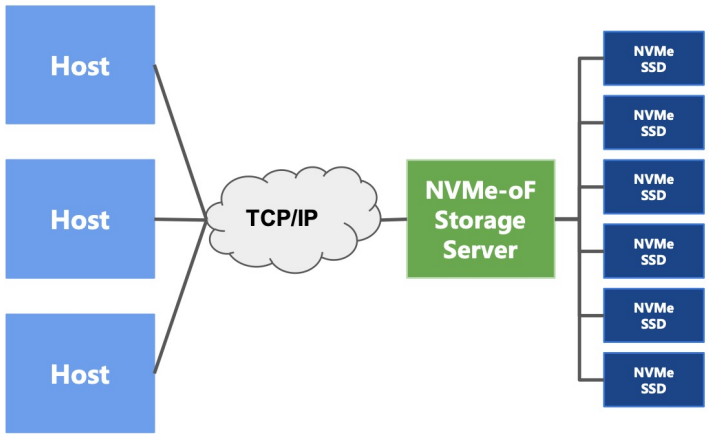
https://nvmexpress.org/

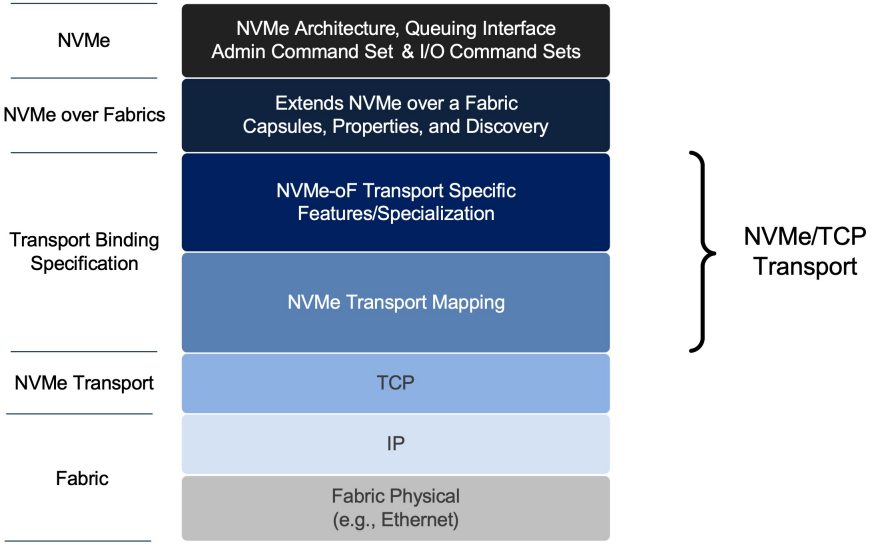
- Non-Volatile Memory express
- Protocolos de transporte y acceso a almacenamiento para SSD
- Conexión directa a CPU a través de PCIe
- Aumenta el paralelismo en las peticiones al disco (64K colas de 64K peticiones cada una)
- Menor latencia
- NVMe-oF : NVMe over Fabrics
  - Uso de DMA (zero-copy)
 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

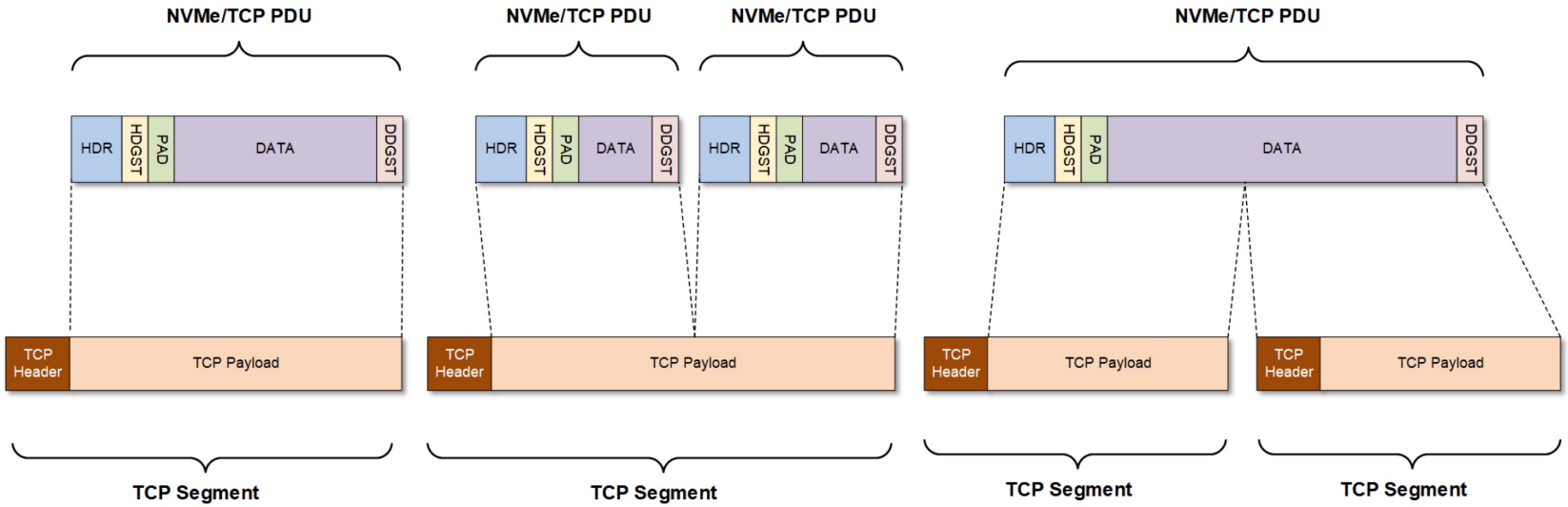


https://nvmexpress.org/wp-content/uploads/NVMe_over_Fabrics_Sept_2017_Brandon_Hoff.pdf

NVMe/TCP







https://nvmexpress.org/wp-content/uploads/March-2019-NVMe-TCP-What-You-Need-to-Know-About-the-Specification.pdf