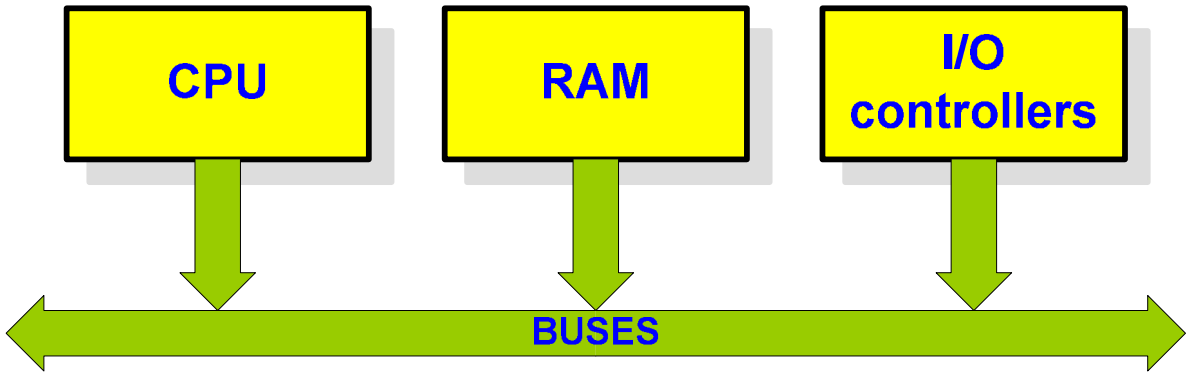
Redes de Almacenamiento

Redes de Interconexión de Periféricos

Administración Avanzada de Sistemas Operativos

Depto. de Arquitectura de Computadores
Universidad de Málaga

Interconexión de la E/S



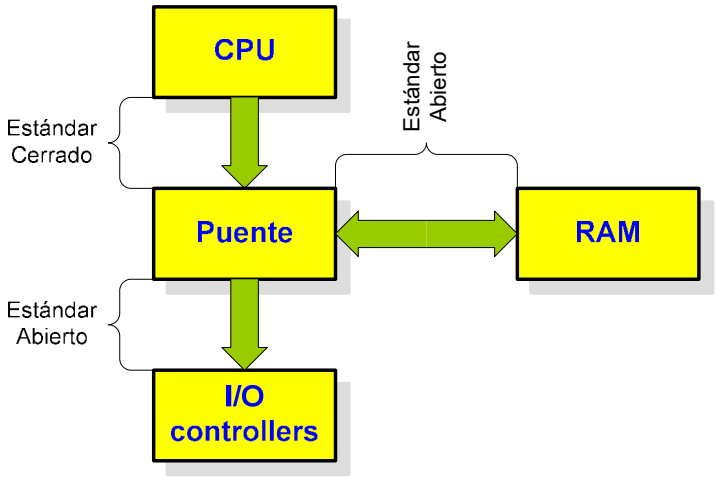
- Arquitectura von Neumann
  - Un bus intercomunica los tres módulos
  - Se usa para transportar datos en forma de “ciclos de transferencia de datos” (direccionamiento+datos+control).
  - En la teoría funciona muy bien, pero…
- …en la práctica real:
  - Si hay varios fabricantes de CPUs y de controladores de I/O se necesita un bus estándar para hacerlos compatibles (un INTERFAZ que oculte la implementación).
  - Debemos distinguir los **fabricantes** de **módulos** (Intel, AMD, Kingston, NVIDIA, ATI, VIA, QLogic, Adaptec, etc), los **integradores** de **sistemas** (IBM, HP, Dell, ASUS, Sony, etc) y los **clientes** finales.

Interconexión de la E/S

- Razones para usar un estándar de interconexión “**abierto**”:
  - Reducir el coste de fabricación del **sistema** final al poder combinar **módulos** de distintos **fabricantes** (competencia y por tanto bajada de precios). El **integrador** puede así ofrecer mejores precios de venta al **cliente**.
  - Cualquier **módulo** nuevo dispone de una amplia biblioteca de módulos compatibles que permiten su **integración** en nuevos diseños inmediatamente. El **integrador** reaprovecha gran parte del diseño.
  - El **fabricante** evita tener que sacar toda una nueva línea de módulos compatibles cada vez que mejora uno de ellos, lo cual reduce la inversión requerida al mejorar un solo módulo.

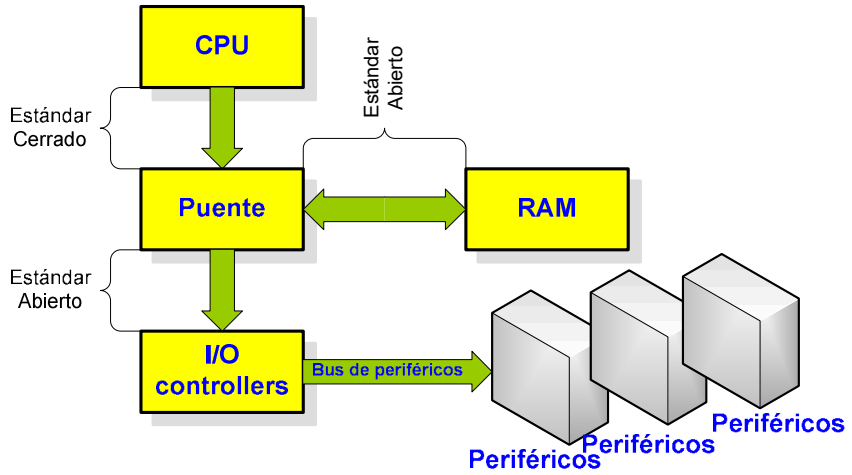
- Razones para usar un estándar “cerrado” (privado al fabricante):
  - Aumentar las ganancias del **fabricante** al reducir la compatibilidad y obligar al integrador a usar todos los módulos del mismo fabricante (menos competencia y por tanto precios más caros).
  - Otros fabricantes pueden usar el estándar cerrado pero pagando una cuota anual por el uso de las **patentes**, con lo cual los productos “clónicos” terminan siendo tan caros como el original. De nuevo gana el **fabricante** evitando la competencia.
  - Esta estrategia es **negativa** a la larga para el fabricante.
    - Los **clientes** prefieren sistemas menos “brillantes” pero más eficientes económicamente (relación precio/rendimiento) que suelen estar fabricados con estándares abiertos.
    - A la larga, la falta de compatibilidad estrangula a la propia empresa debido a que debe costear el diseño de todos los módulos cada vez que introduce una nueva tecnología. (Por ejemplo, el caso de DEC con el diseño del procesador Alpha. DEC quebró cuando no puedo vender suficientes procesadores como para costear el diseño del siguiente).

Interconexión de la E/S



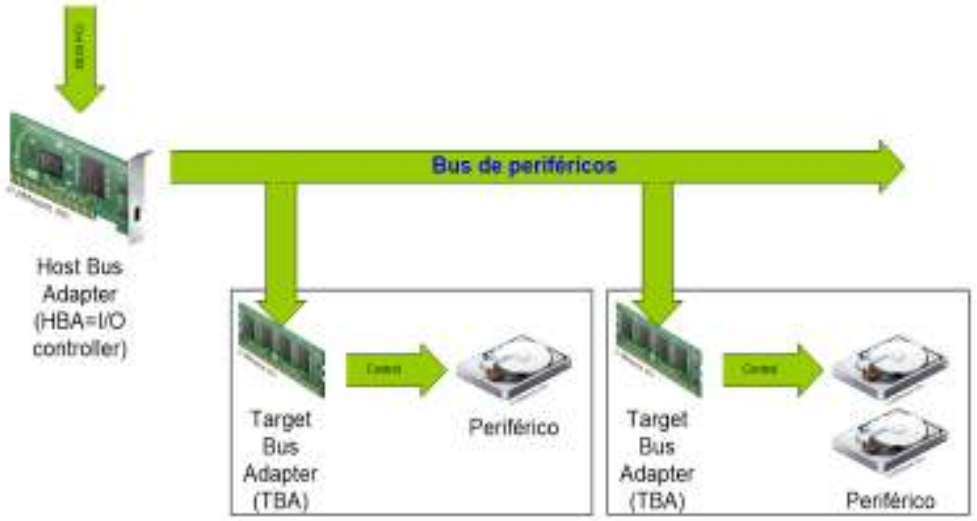
- La solución real pasa por usar un adaptador (**puente**) entre el bus del fabricante de la CPU y los buses hacia el resto de los módulos.
  - Bus del fabricante: Front Side Bus, HyperTransport, QuickPath Interconnect
  - Bus de memoria: DDR
  - Bus de controladores de I/O: PCI, PCI-X, PCI-express, InfiniBand, HyperTransport, etc..
- PCI y sus evoluciones no son los mejores, pero están definidos por un comité abierto que **no cobra por el uso de las patentes**.
  - Cualquier fabricante puede implementar un controlador con interfaz PCI sin costes adicionales por las patentes.
  - La oferta de módulos disponibles para el bus PCI, PCI-X o PCI-e es enorme

Bus de periféricos



- El bus de periféricos está diseñado para lograr unos objetivos distintos a otros buses de la máquina:
  - Conectar varios periféricos a un solo controlador (decenas de ellos).
  - Cubrir distancias de varios metros (e incluso kilómetros) para llegar a periféricos voluminosos (p.e. RAIDs o librerías de cinta).
- La tecnología del bus de periféricos es diferente a la del bus de controladores de I/O al tener objetivos distintos.

Bus de periféricos: términos

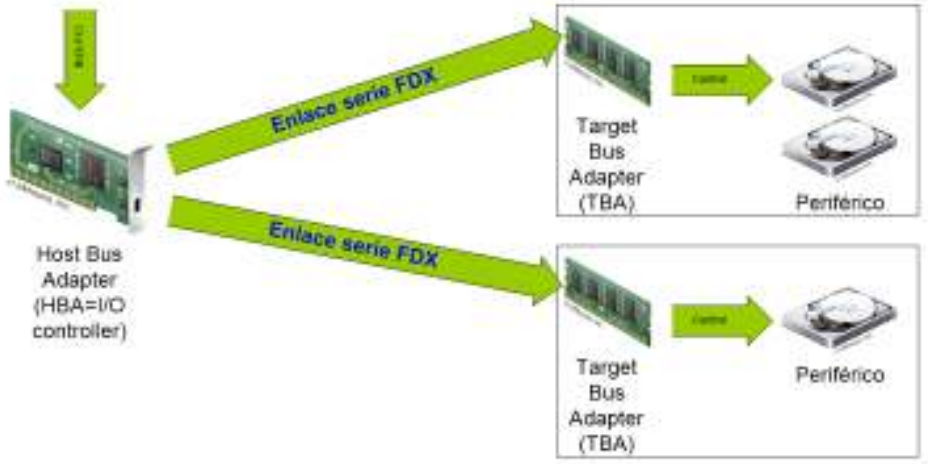


- Host Bus Adapter (HBA): Puente entre el bus de controladores (PCI normalmente) y el bus de periféricos (ATA, SATA, SCSI, SAS, FC, etc).
- Target: Es el conjunto de un periférico más la lógica que lo conecta al bus (TBA). La implementación interna del target es totalmente invisible al HBA.
- Target Bus Adapter: Lógica de un target que se ocupa de gestionar las comunicaciones con el HBA y de ejecutar las operaciones sobre el periférico.

Evolución de los buses

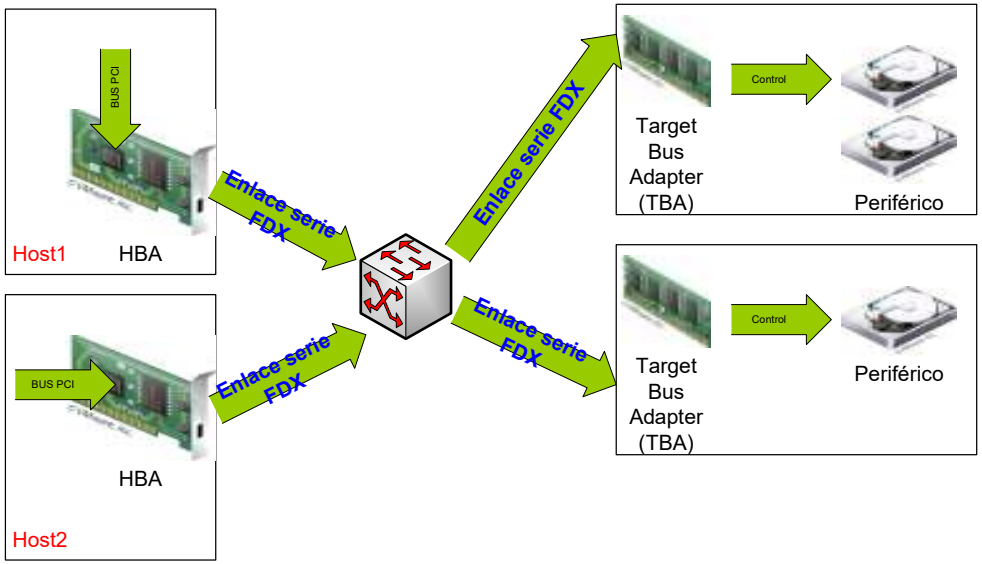
- Serialización:
  - Es más difícil aumentar la frecuencia de un bus con líneas paralelas que de uno con una sola línea de datos en serie.
    - PCI Express: Serialización de PCI
    - SATA: Serialización de ATA/IDE
    - SAS: Serialización del nivel físico de SCSI
    - FC: Diseñado como bus serie
  - Los buses serie pueden usar varios enlaces en paralelo además, pero no están sincronizados entre si.
- Topología
  - Los buses paralelos suelen ser "multidrop« (bus físico con varios HBA y TBA conectados y un protocolo de arbitraje) y Half-Duplex. El ancho de banda es compartido si se transfieren datos simultáneamente a varios Targets.
  - Los buses serie suelen usar un HBA que implementa un switch y enlaces punto a punto Full-Duplex con los TBA. No hay arbitraje ni compartición de ancho de banda entre el HBA y cada Target.

Bus de periféricos serie



- Host Bus Adapter (HBA): Implementa una topología en estrella con enlaces privados.
- Target Bus Adapter (TBA): Dialoga punto a punto con el HBA.
- Estándares para la transferencias de datos por el bus de periféricos:
  - SCSI, SATA, FC

Red de periféricos (SAN)



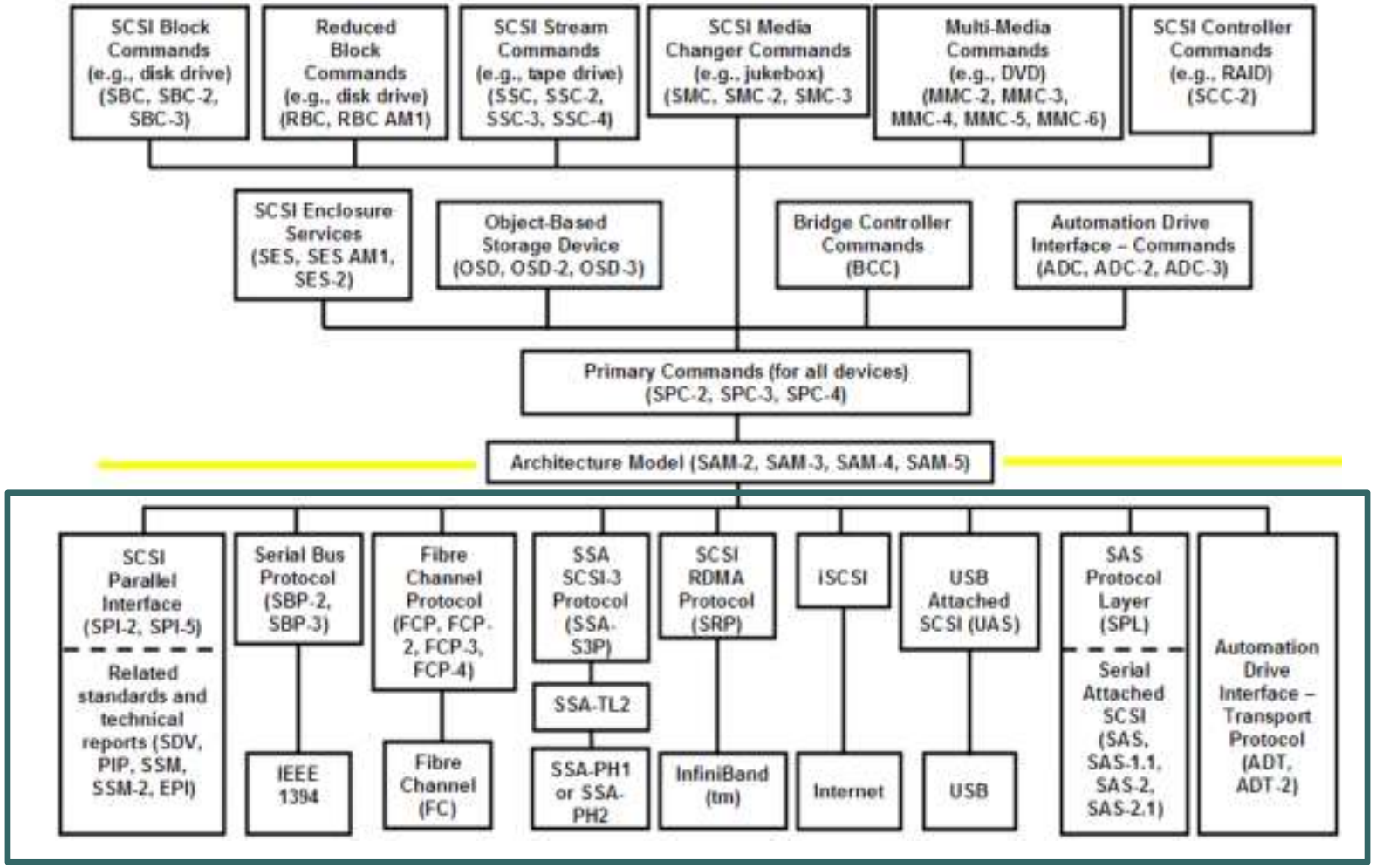
- El objetivo es compartir periféricos entre varios sistemas distintos.
- Se intercala un switch entre los HBAs y los TBAs.
- El bus ha sido diseñado para funcionar con conmutación de paquetes:
  - Nivel físico con enlaces serie y Full-Duplex
  - Nivel de enlace basado en paquetes

Arquitectura SCSI

- Estándar definido para el bus de periféricos con capas independientes similares a OSI
- 3 capas:
  - Nivel físico (SCSI Interconnect Layer): múltiples soluciones para escenarios distintos y permitir la evolución tecnológica
    - Bus SCSI paralelo original
    - SAS (Serial Attached SCSI): Enlace serie en cobre
    - Fibre Channel: Red óptica
    - FCoE (Fibre Channel on Ethernet), FCIP and iFCP (Fibre Channel on IP)
    - iSCSI (SCSI on TCP)
  - Nivel de enlace (SCSI Transport Protocol Layer): Estructura la comunicación y el diálogo entre HBAs y TBAs en forma de paquetes de datos.
  - Nivel de comandos (SCSI Application Layer): Define comandos comunes a todos los periféricos y comandos específicos a cada tipo (discos, cintas, librerías de cinta, etc).
- Los niveles medio y alto se han mantenido casi inmutables mientras que en el nivel físico ha habido y sigue habiendo una gran evolución.
  - Todos los periféricos SCSI dialogan con los mismos comandos independientemente de su conexión.
  - La compatibilidad entre fabricantes de HBAs y periféricos, y a lo largo del tiempo está garantizada.

Arquitectura SCSI

Alternativas para el transporte de comandos SCSI

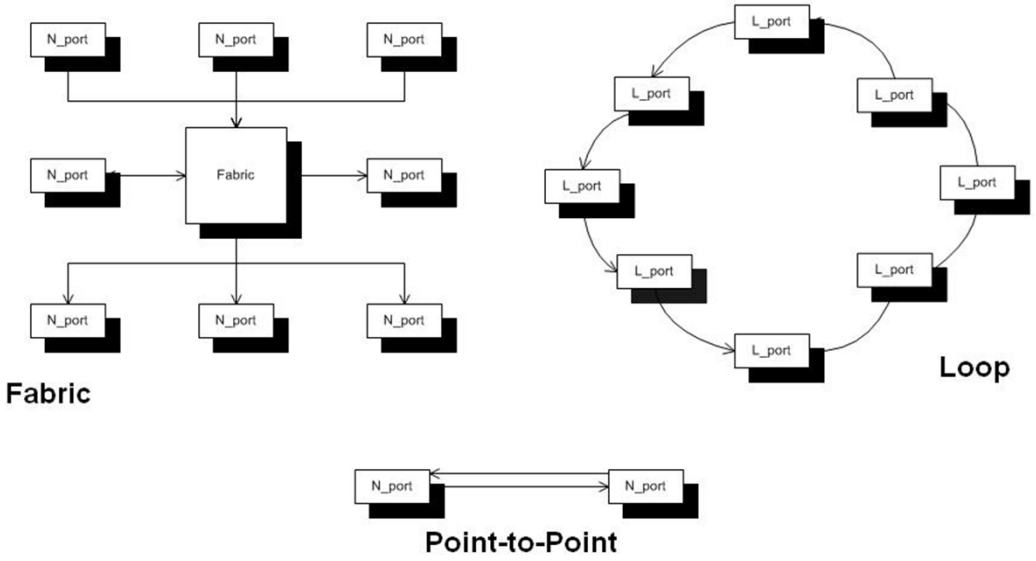


Arquitectura SATA

- El estándar para el bus de periféricos ATA/SATA también define capas.
  - La capa física es tan parecida a SAS que los periféricos SAS son capaces de reconfigurarse para funcionar en un bus SATA.
  - Las capas de enlace y comandos son distintas a SCSI aunque similares en filosofía.
  - Las limitaciones del estándar SATA hacen más fácil la implementación del TBA y por tanto más barata, reduciendo el precio de los periféricos.

Fiberchannel

- Capa de protocolos que permite el transporte de comandos SCSI.
- Topologías
  - Conexión punto a punto
  - Bucle arbitrado
  - Red conmutada (el más usado en arquitecturas de almacenamiento distribuido)



Fibre Channel: capas

FC-3: encriptación o algoritmos de redundancia RAID

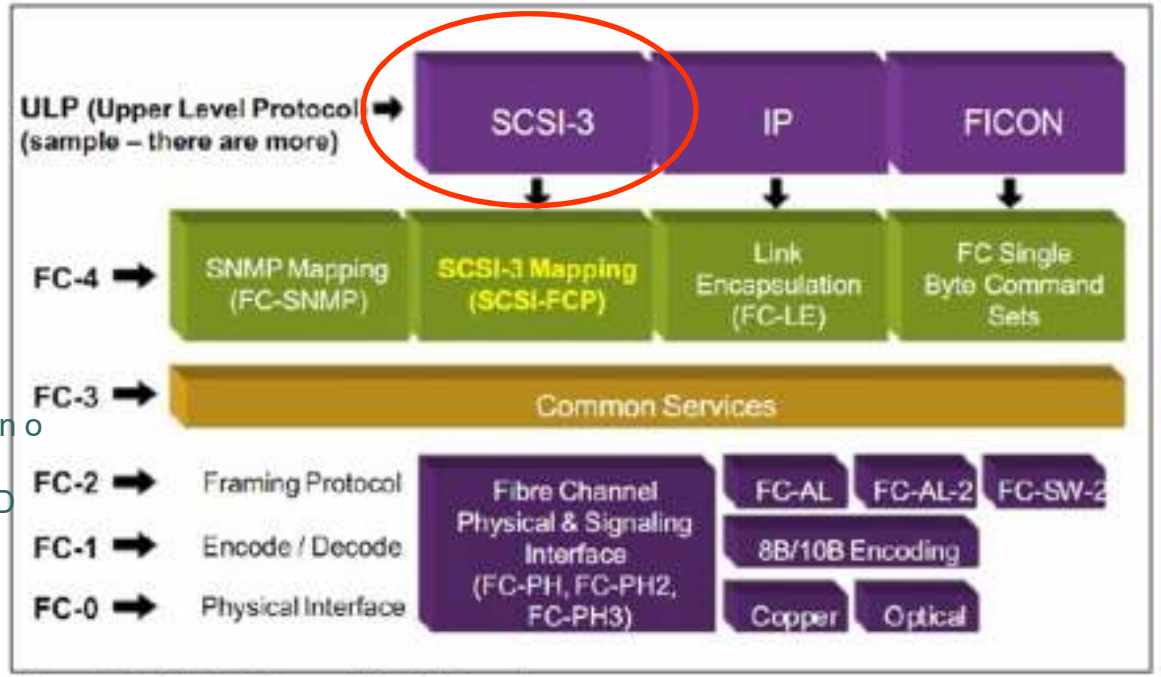
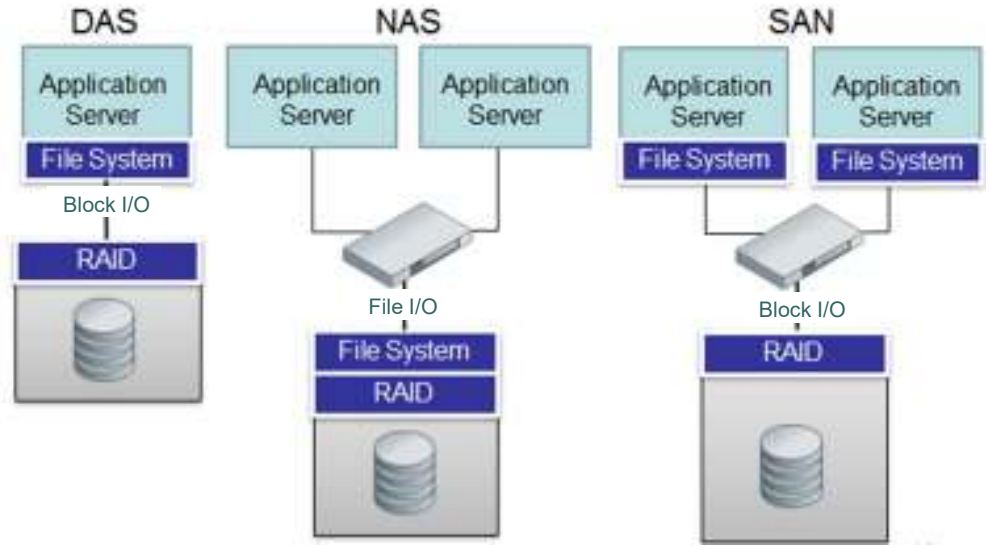


Figure 2-4 Fibre Channel Model Overview

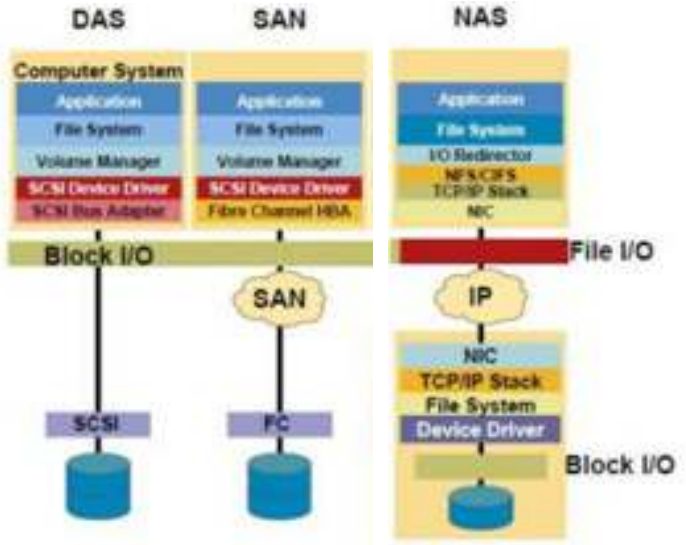
Modelos de Almacenamiento



- 3 modelos de almacenamiento según la topología
  - DAS (Direct Attached Storage): 1 periférico solo se conecta a 1 máquina (modelo tradicional en desktop y laptop). Los datos solo son accesibles en un sistema.
  - NAS (Network Attached Storage): Se usa un sistema DAS y se conecta a una LAN. Se usa un sistema de ficheros en red para acceder a los datos del servidor desde otros clientes. Es menos eficiente ya que los sistemas de ficheros en red tienen más overhead que un bus de periféricos. Típicamente aplicación sobre la torre TC/IP
  - SAN (Storage Area Network): 1 periférico está conectado a más de un sistema a través de una red de almacenamiento. Cualquier sistema puede acceder a los datos con a velocidad de transferencia nativa del bus de periféricos. Red sobre fiberchannel, pero también tecnologías alternativas (Ethernet, IP).

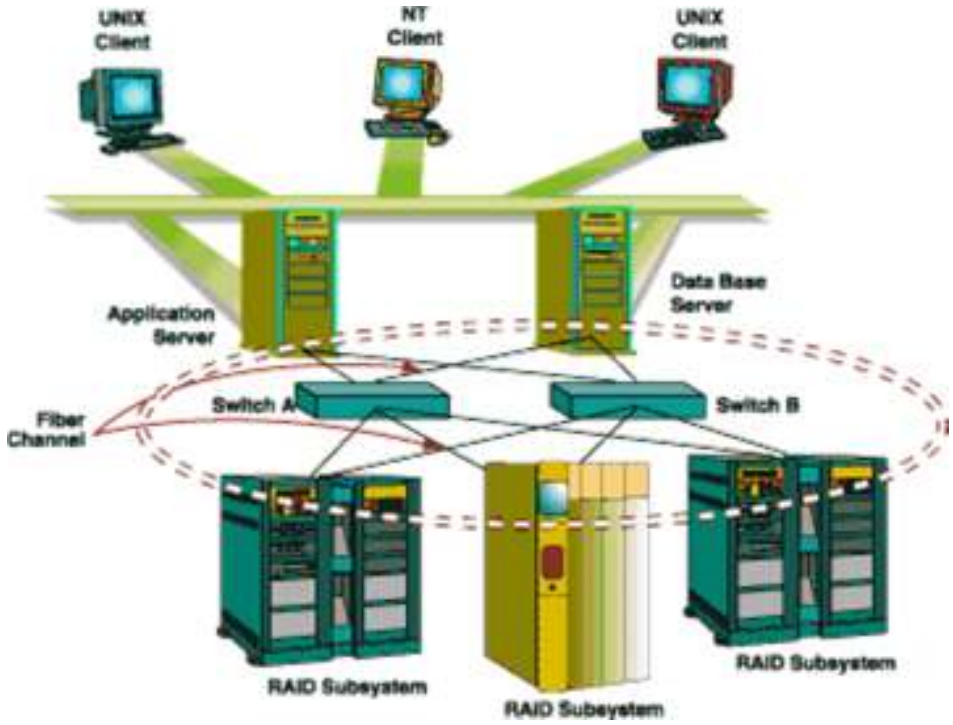
Acceso Block I/O vs File I/O

- Servidores DAS y SAN dialogan directamente con los dispositivos usando Block/IO
  - Comandos SCSI
- NAS usa File I/O
  - Llamadas al sistema de ficheros



SAN con fiberchannel

- Diseño con redundancia



Dos tipos de switches:

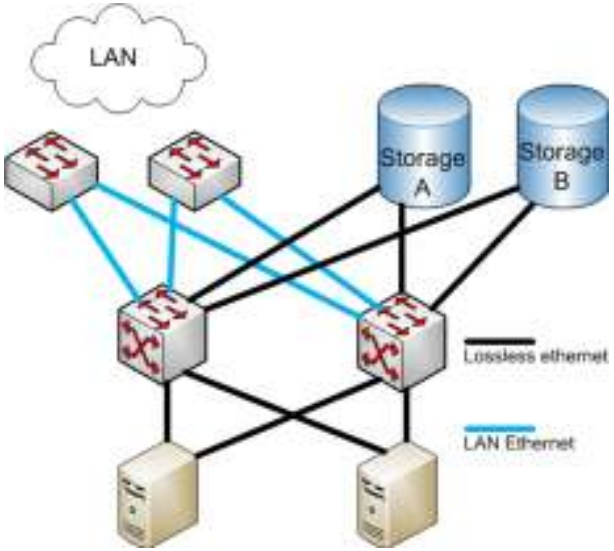
Directors: alto número de puertos y alta disponibilidad (redundancia de fuente módulos de control, reemplazo en caliente, …)

Switches: menos puertos, menos redundancia

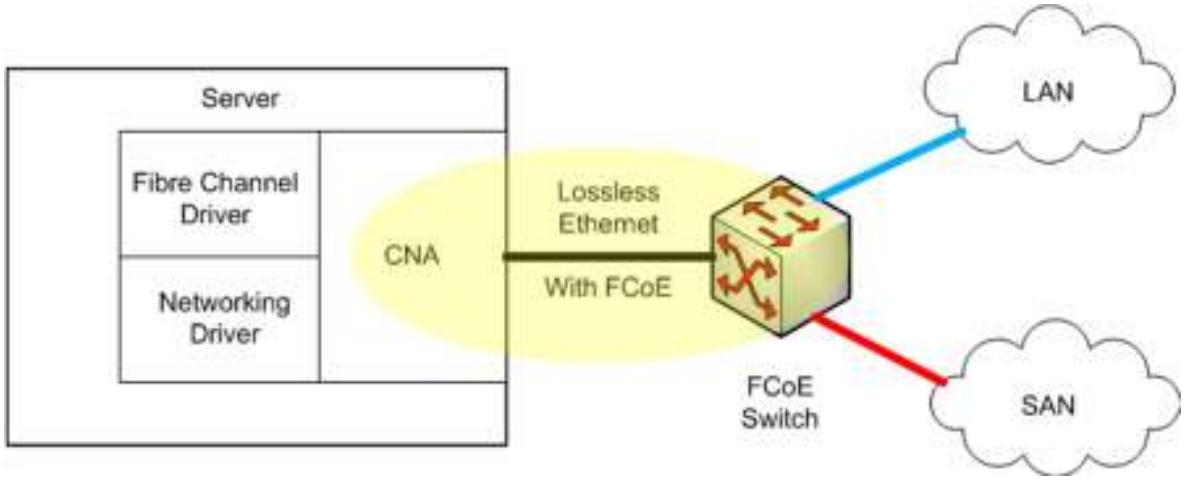
Componentes Fibre Channel

SAN con FCoE

- FiberChannel over Ethernet encapsula comandos FiberChannel sobre la capa Ethernet
  - Así se pueden usar switches Ethernet para transportar Fiberchannel
  - Capa FC1 y FC2 son sustituidas por capas Ethernet
    - Control de flujo basado en prioridades debe ser añadido a los switches ethernet (más caros).



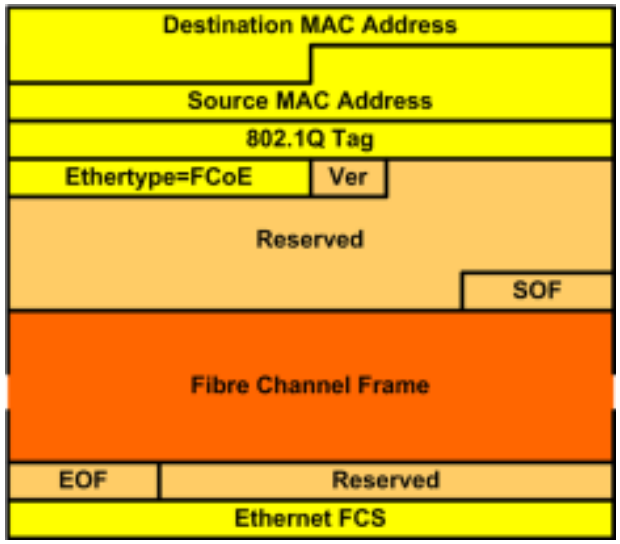
FCoE



- Permite la convergencia de la red
  - Distintos protocolos de interconexión comparten la misma red: Aplicaciones sobre TCP/IP y comandos FC para almacenamiento conviven en la misma red
    - Puede ser recomendable el uso de VLANS para dividir tráfico
  - CNA (Converged Network Adapters)
    - Contiene HBA de Fiber Channel y NIC Ethernet reduciendo el número de tarjetas, cables, switches y consumo de energía.

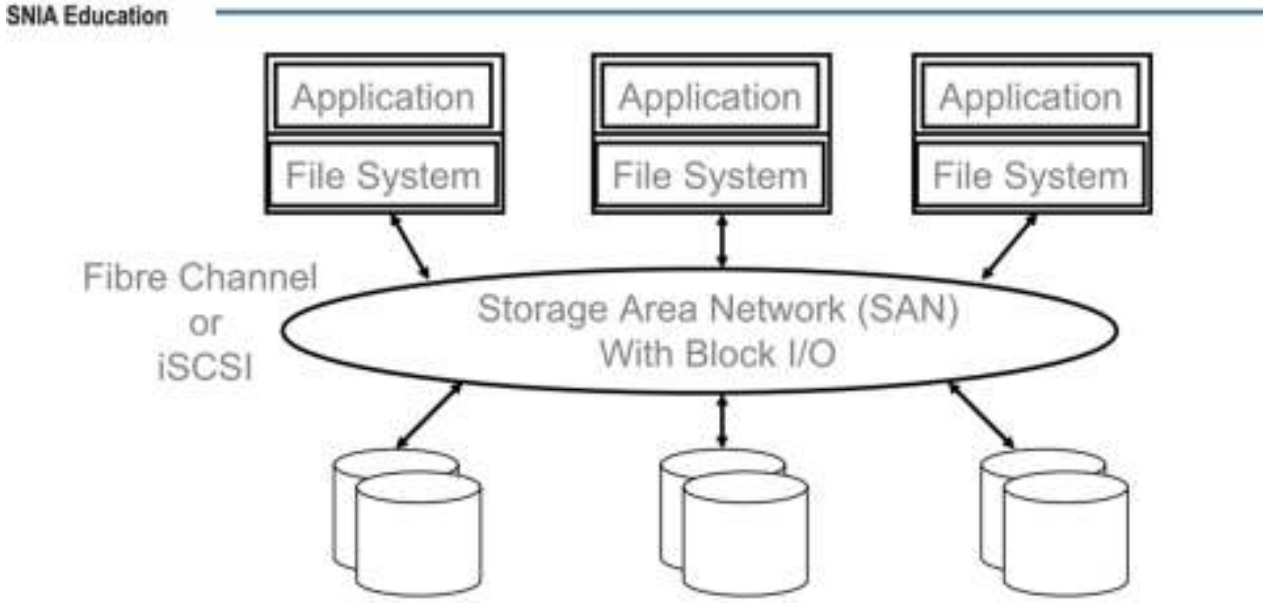
FCoE

- Trama FCoE
  - Correspondencia entre direcciones MAC y puertos FC

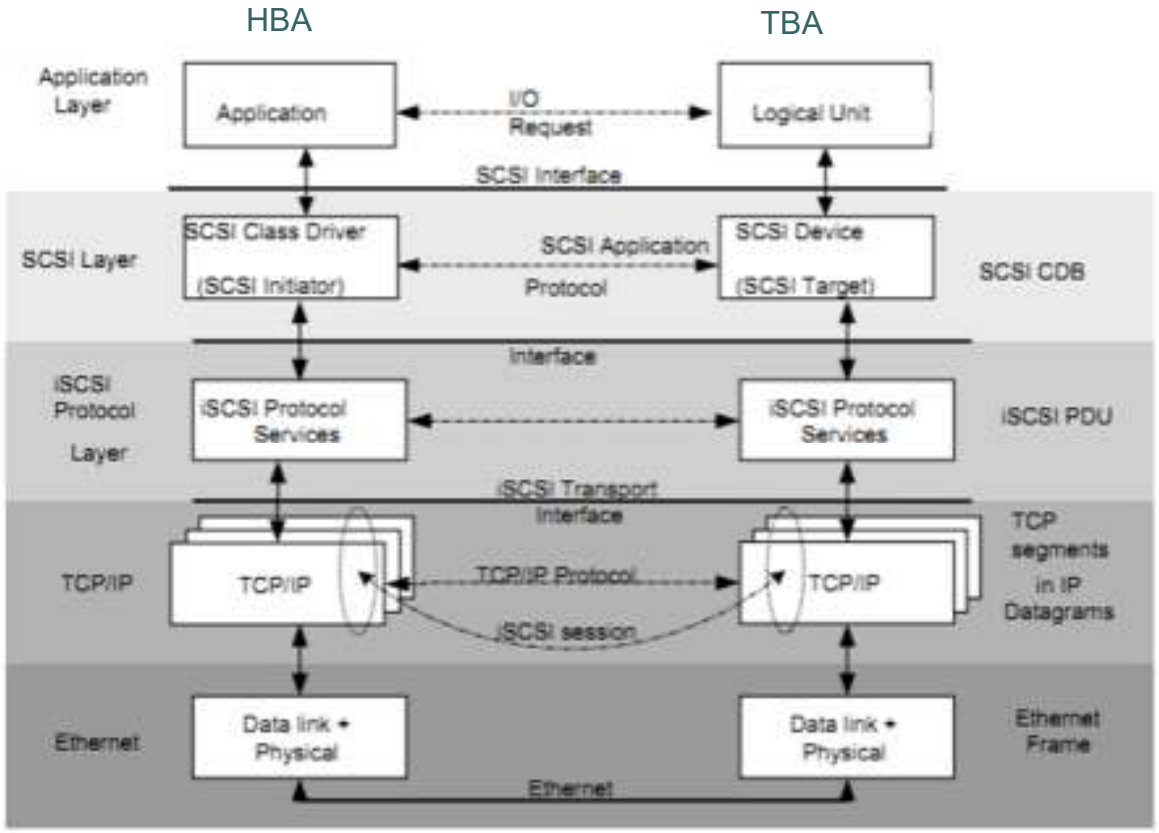


SAN con iSCSI

- SCSI sobre IP
- iSCSI puede implementar una SAN
  - El switch podrá ser Ethernet
  - Routers también puede ser usados

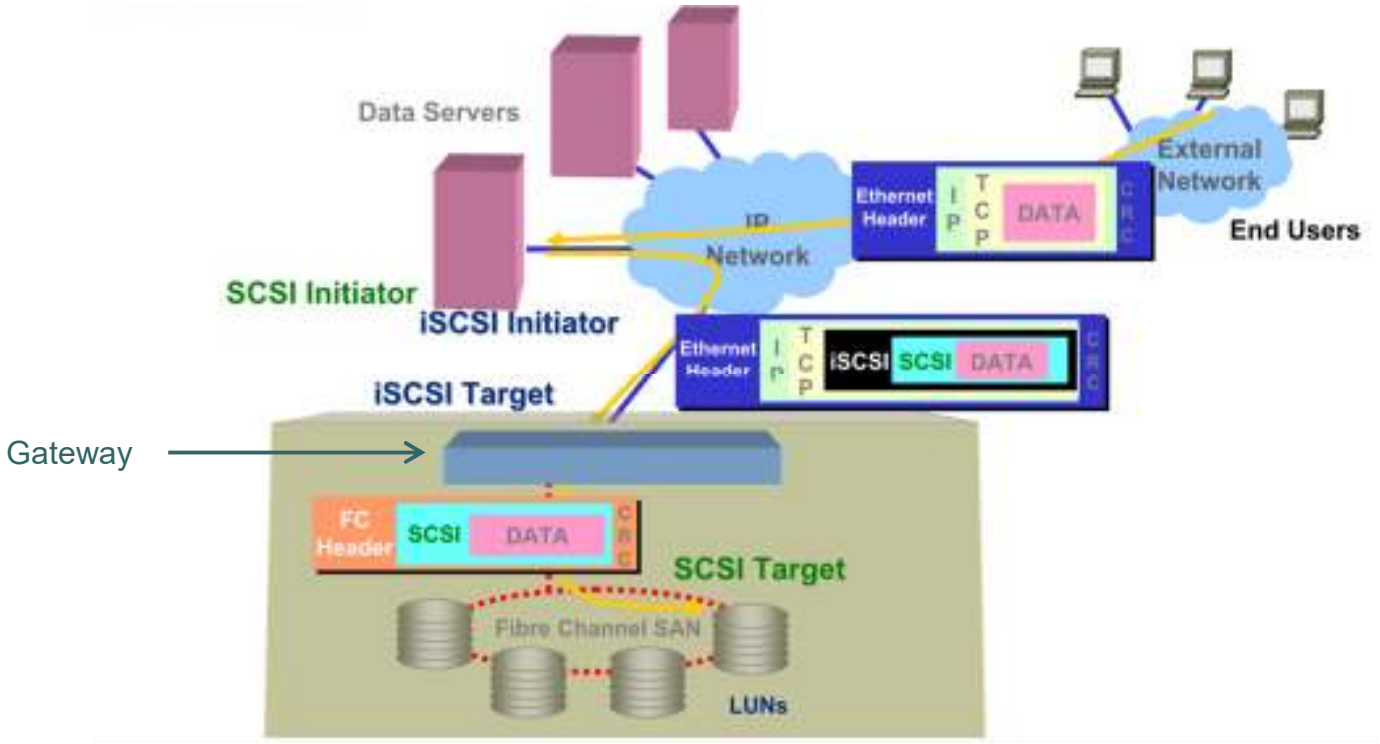


Torre de protocolos iSCSI



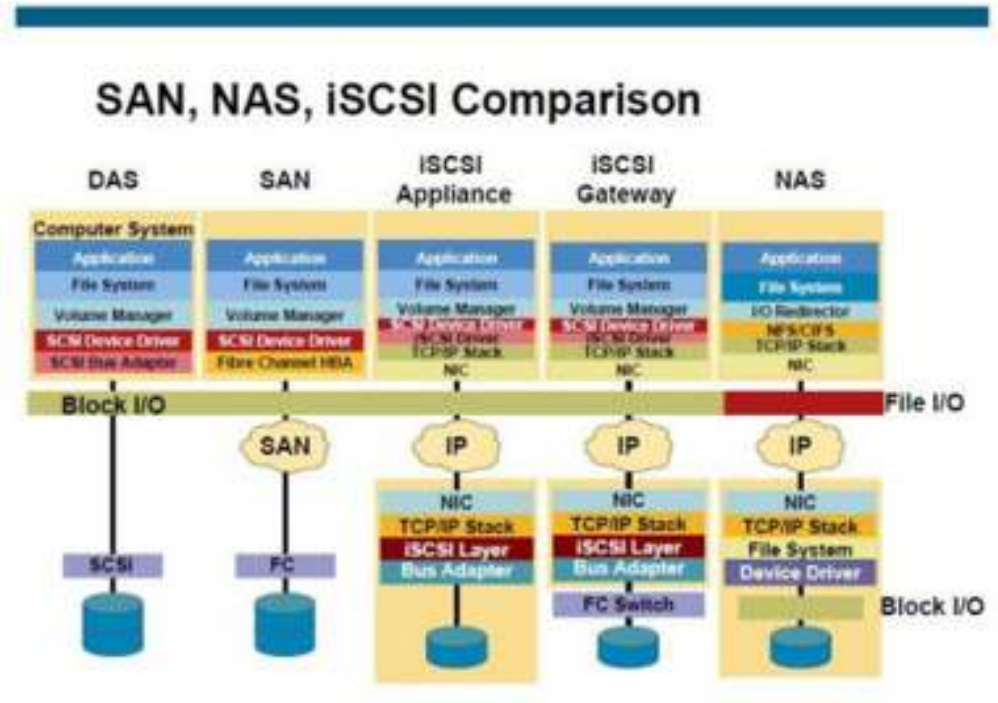
iSCSI para acceso externo

- End user accede a un equipo de la intranet que inicia acceso a una SAN usando iSCSI



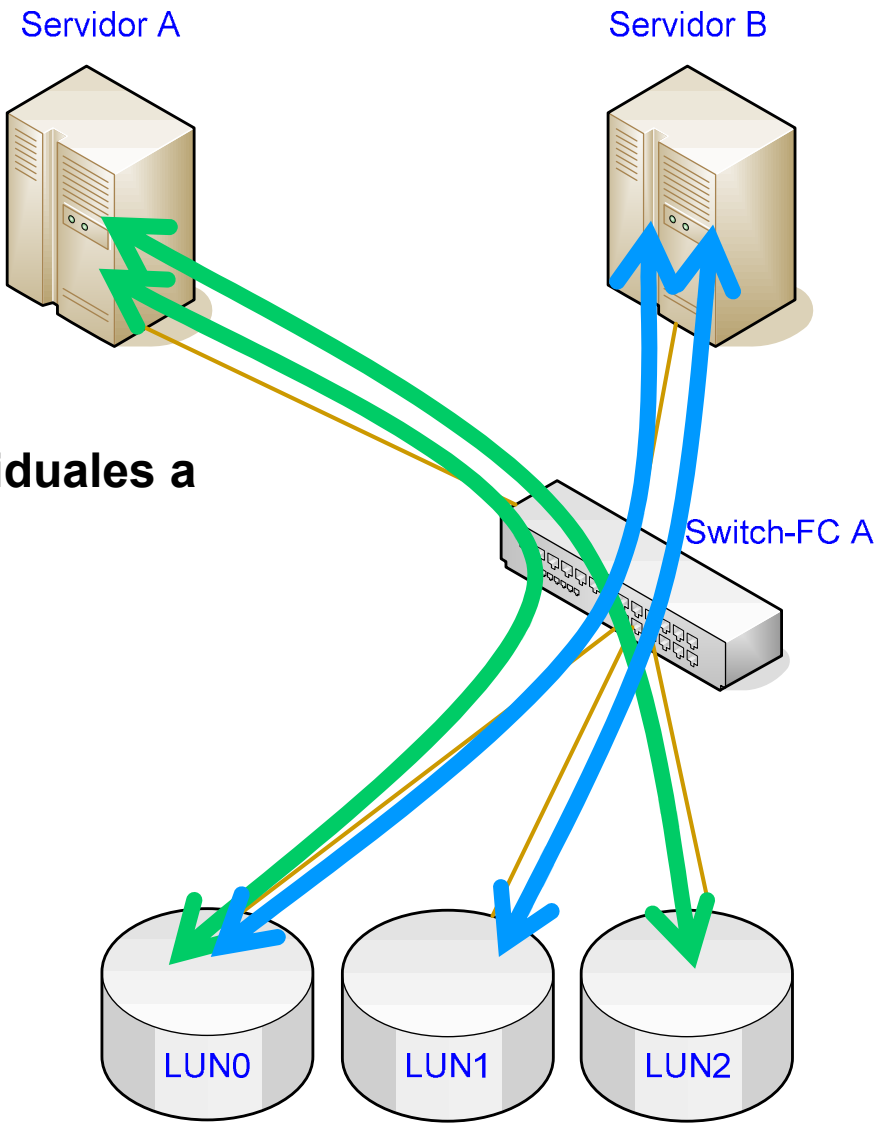
Adaptadores iSCSI (HBAs)

- Tarjetas de red que incorporan funcionalidad de proceso iSCSI integrada.
  - HBA iSCSI son tratado por el SO como controladores SCSI convencionales

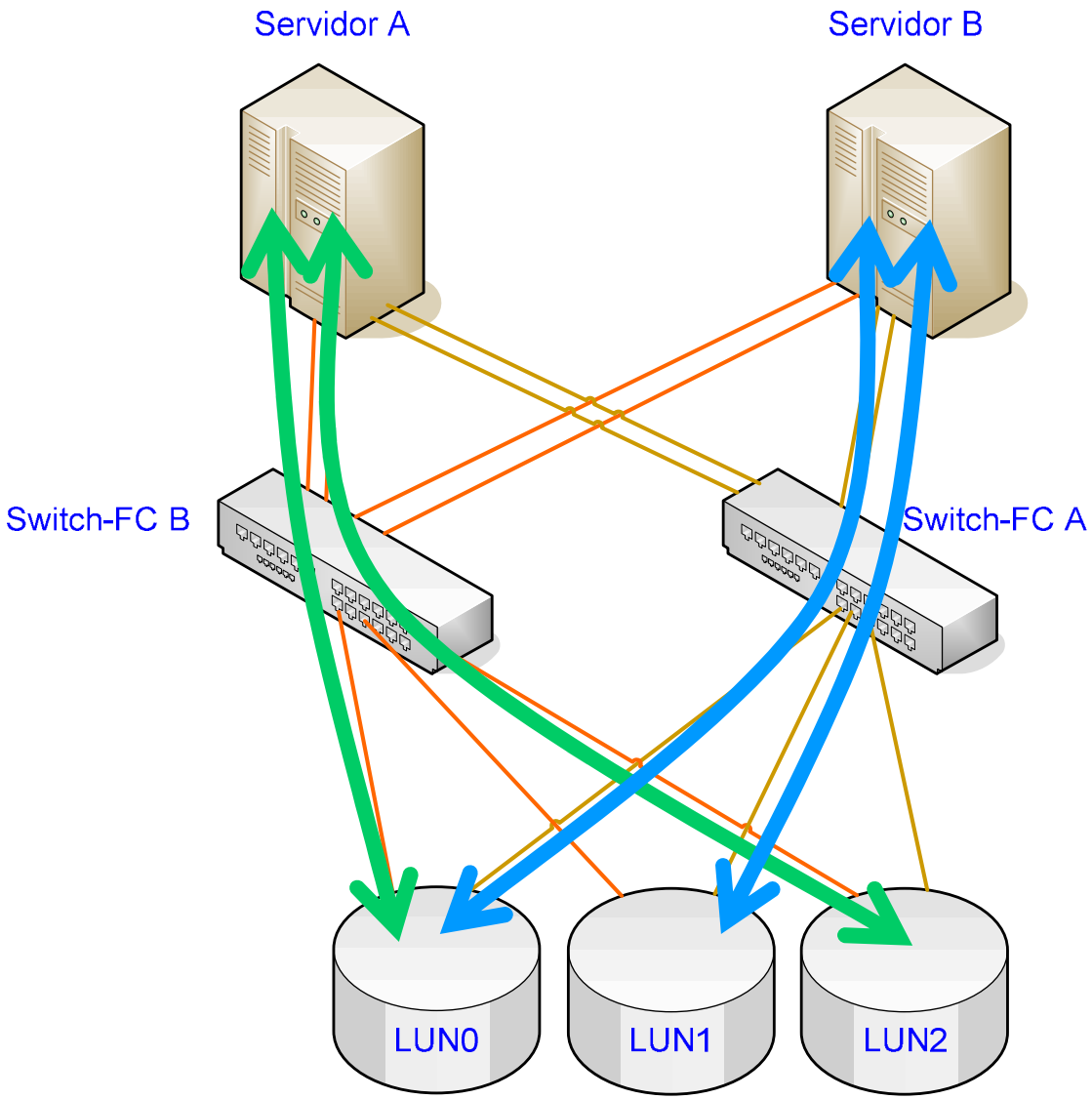


Infraestructura Básica SAN

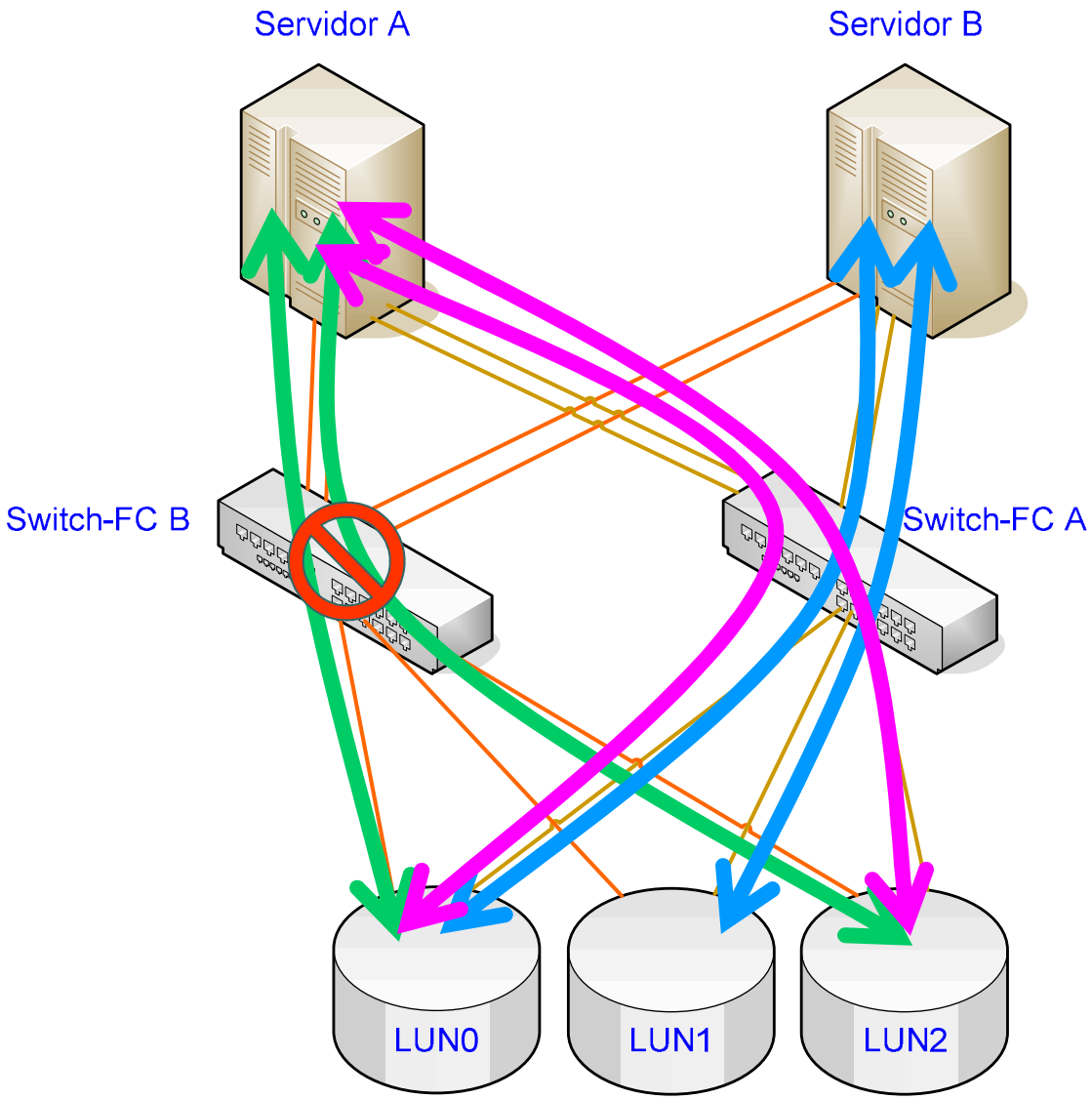
Enlaces individuales a
- **2Gbps**
- **4Gbps**
- **8Gbps**



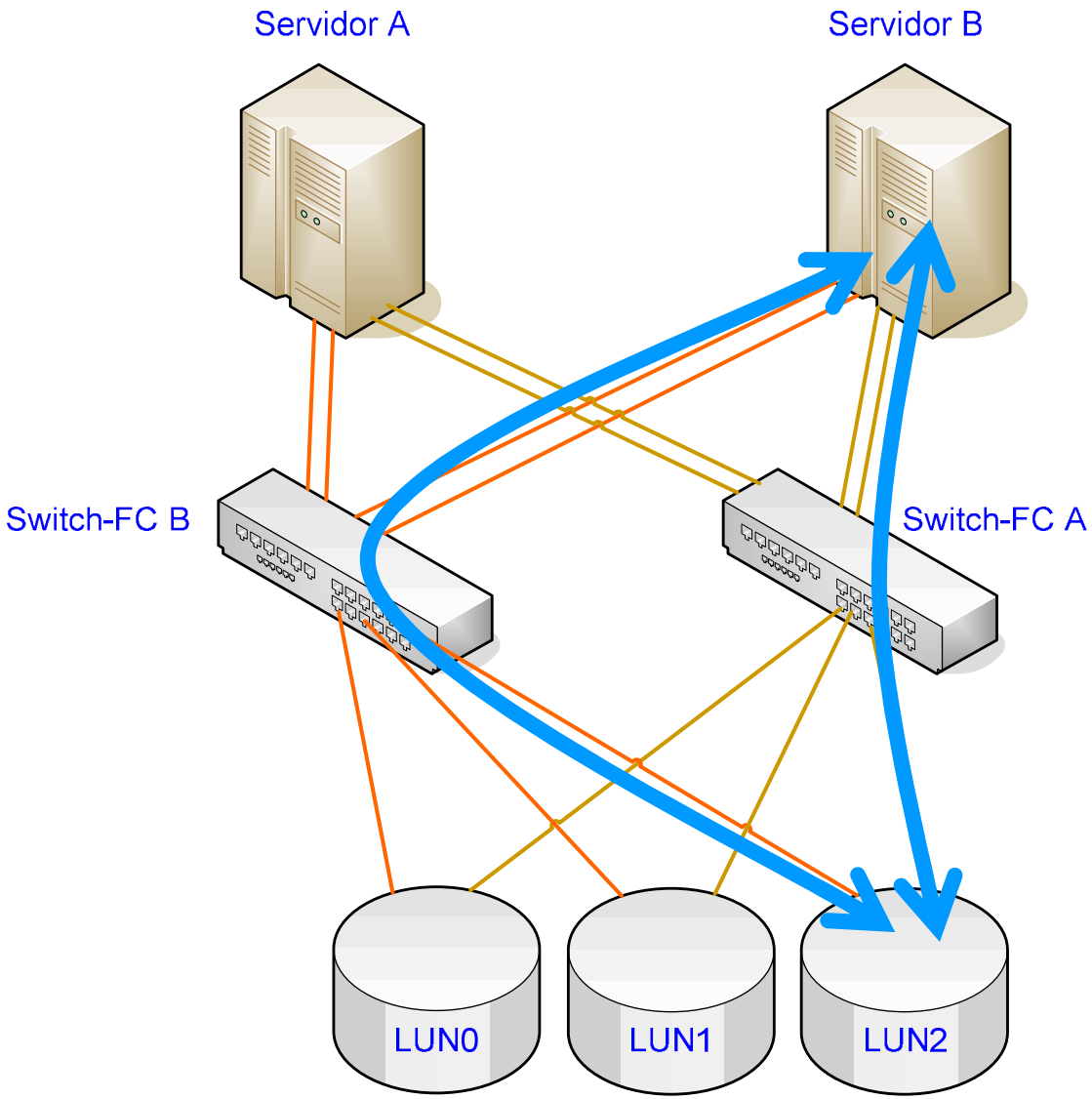
Infraestructura Redundante SAN



Recuperación Automática



Agregación de Ancho de Banda



Redundancia hardware

- Servidores redundantes
  - Fuente de alimentación duplicada
  - Ventiladores replicados
  - Varias CPUs
  - Memoria RAM con corrección de errores (ECC)
  - Componentes sustituibles en caliente (hotplug)
  - **Discos de sistema en espejo (RAID1)**
  - **Controladora de disco replicada**
  - Interfaces de red duplicados y agregados con protocolo LACP
  - Switches de red duplicados

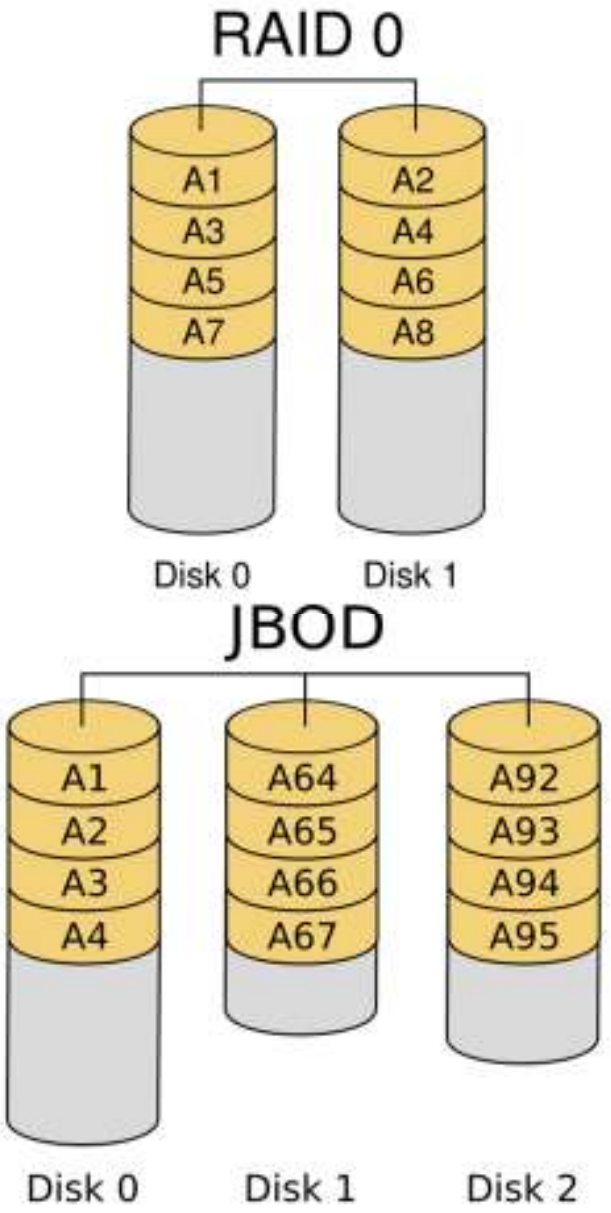
- Almacenamiento redundante
  - **Almacenamiento de disco con controladora en modo RAID1, 5 o 6**
  - **Controladoras de disco RAID redundantes**
  - **Enlace de comunicaciones redundante conectando el RAID a dos switches SAN diferentes**
  - **Enlace de comunicaciones redundante entre cada servidor y los switches SAN**

Redundancia software

- Sistema Operativo
  - Hot-plugin: Permite cambiar componentes hardware sin apagar
  - Driver de red con soporte LACP (enlaces de red redundantes)
  - **Driver de SAN multipath (gestiona caminos redundantes hasta el disco)**
  - **Sistema de ficheros en cluster (permite montar la misma partición en varias máquinas)**
- Migración y distribución
  - Migración de aplicaciones entre servidores en caso de avería de uno
  - Ejecución distribuida simultánea entre varios servidores
  - Posible distribución de datos entre CPDs
- Switches de red
  - Soporte *split multi-link trunking* para permitir varios caminos al mismo destino en la red
  - Switches redundantes

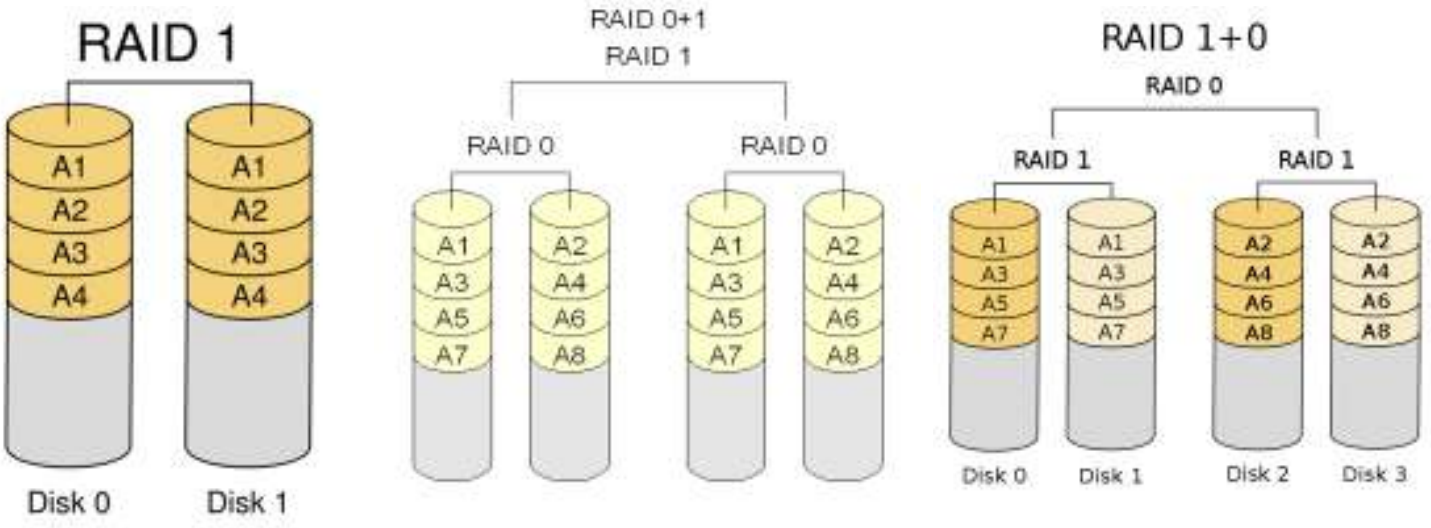
RAID0 y JBOD

- Este modo concatena los discos para aumentar la capacidad sin ningún tipo de redundancia.
- La diferencia entre ambos modos es la distribución de los bloques.
- En RAID0 hay paralelismo entre discos cuando se acceden secuencialmente varios bloques, aumentando la velocidad de transferencia.
- En caso de avería de 1 disco se pierde el contenido completo del disco lógico.



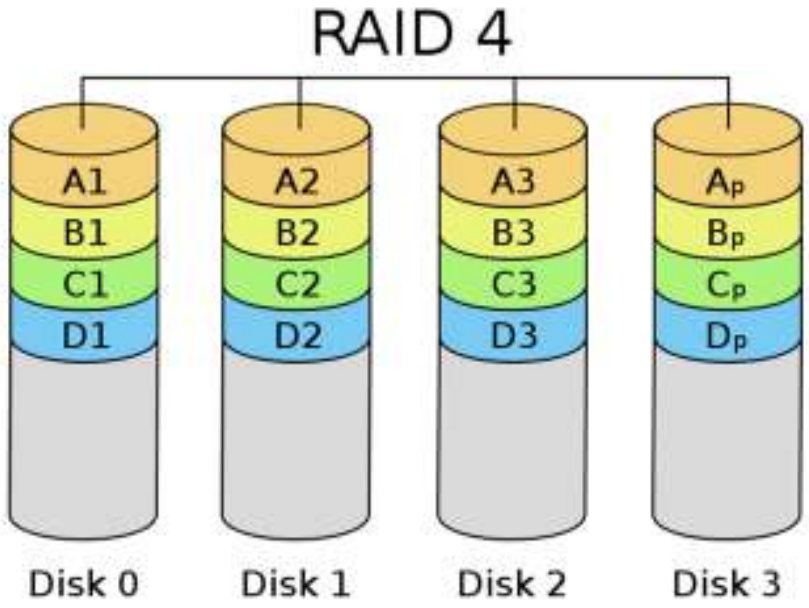
RAID1

- Se usa un número par de discos. Cada bloque del volumen lógico está copiado en dos discos físicos distintos (mirroring).
- En caso de fallo de un disco, no se para el acceso al volumen lógico y se continúa leyendo y escribiendo en el disco restante (modo degradado). El fallo de un segundo disco es catastrófico ya que se pierden los datos.
- Al reemplazar el disco averiado, el controlador reconstruye su contenido copiando el del otro disco mientras continúa funcionando normalmente.
- Normalmente la reconstrucción es automática y se definen políticas de prioridad entre los accesos de las aplicaciones y los del controlador RAID durante la reconstrucción.
- También existen combinaciones RAID0+1 y RAID1+0 que determinan distribuciones de los bloques de datos entre varios discos para sacar partidor del mejor rendimiento del RAID0.



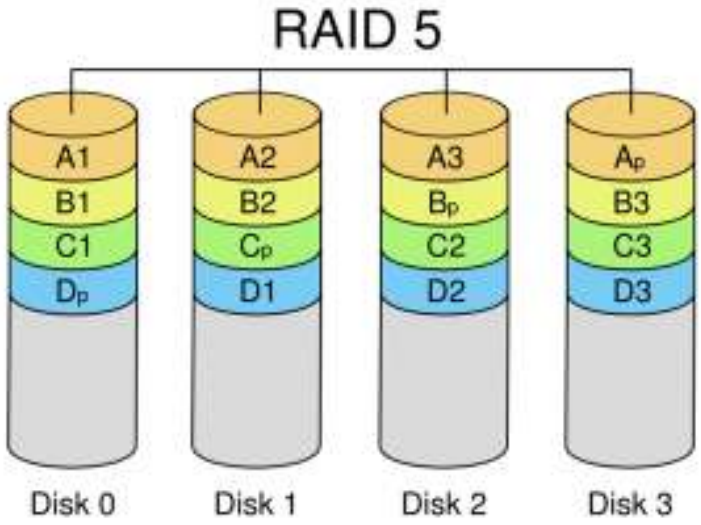
RAID4

- Los bloques consecutivos se reparten entre discos distintos para paralelizar las lecturas y escrituras de secuencias de información y conseguir mayor velocidad de transferencia de datos.
- La redundancia para n discos se consigue calculando un bloque con la paridad de los n-1 bloques consecutivos guardados en discos distintos y guardándolo en un disco dedicado a paridad.
- En caso de avería de un disco, se continúa escribiendo normalmente en los discos restantes. En las lecturas se leen los datos de los discos restantes y se calcula el contenido del disco ausente gracias a la paridad (modo degradado).
- Al cambiar el disco averiado, se reconstruye su contenido automáticamente a partir de la paridad.



RAID5/6

- Similar a RAID4, pero el bloque de paridad va rotando de disco para evitar convertir el disco de paridad en un cuello de botella en el caso de un patrón aleatorio de pequeñas escrituras.
- Tanto RAID4 como RAID5 no admiten más de un disco averiado a la vez. El RAID6 es un RAID5 con 2 discos extra para paridad para sobrevivir a 2 averías de discos seguidas.



Comparativa RAIDs

- RAID 0: Mayor velocidad de transferencia en patrones secuencial y random, ninguna redundancia.
- RAID 1: Tolerancia a 1 fallo. Ninguna mejora de rendimiento.
- RAID 4/5: Tolerancia a 1 fallo. Mayor velocidad de transferencia en patrones secuenciales y random de lectura. En patrones random de escritura se ralentiza porque hay que leer el resto de los discos para recalcular la paridad de una fila de bloques.
- RAID 6:Tolerancia a 2 fallos. Eficiencia similar a RAID5.
- RAID0+1: Es como un RAID0 replicado en mirror sobre otro RAID0. No tolera fallos de dos discos si no están en el mismo mirror. Combina el rendimiento del RAID0 con la redundancia del RAID1.
- RAID1+0: Es un RAID0 compuesto por dos RAID1. Es algo más tolerante a fallos de discos que el RAID0+1 con el mismo rendimiento.