Tema 7. Sistemas de almacenamiento

ESTRUCTURA DE COMPUTADORES

Grupo de Arquitectura de Computadores (GAC)

Objetivos del tema

En el curso pasado (**Fundamentos de Computadores**) se estudiaron algunos conceptos relativos al manejo de la Entrada/Salida de un computador, como la gestión de operaciones de E/S mediante E/S programada, E/S con interrupciones o E/S con DMA.
En este curso se profundizará en el estudio del subsistema de E/S, prestando especial atención a los dispositivos de almacenamiento y al sistema de interconexión mediante buses.

En este tema se estudiará:

- Las principales opciones de almacenamiento secundario de un sistema
- Qué es y cómo configurar un array de discos (RAID)

Bibliografía

Básica

- *Computer Organization and Design: The hardware/software interface (3rd ed.)*. David A. Patterson and John L. Hennessy. Morgan Kaufmann Publishers, Inc. 2007
- *Computer Architecture: A Quantitative Approach (3rd or 4th ed.)*. John L. Hennessy y David A. Patterson. Morgan Kaufmann Publishers, Inc. 2004/2006
- *Organización y arquitectura de computadores (7th ed.)*. William Stallings. Prentice Hall. 2006

Complementaria

- *Organización de Computadores*. C. Hamacher, Z. Vranesic y S. Zaky. Mc Graw Hill. 2003
- *Problemas Resueltos de Estructura de Computadores*. F. García, J. Carretero, J.D. García y D. Expósito. Paraninfo, 2009.

Índice

Índice

La E/S en un computador

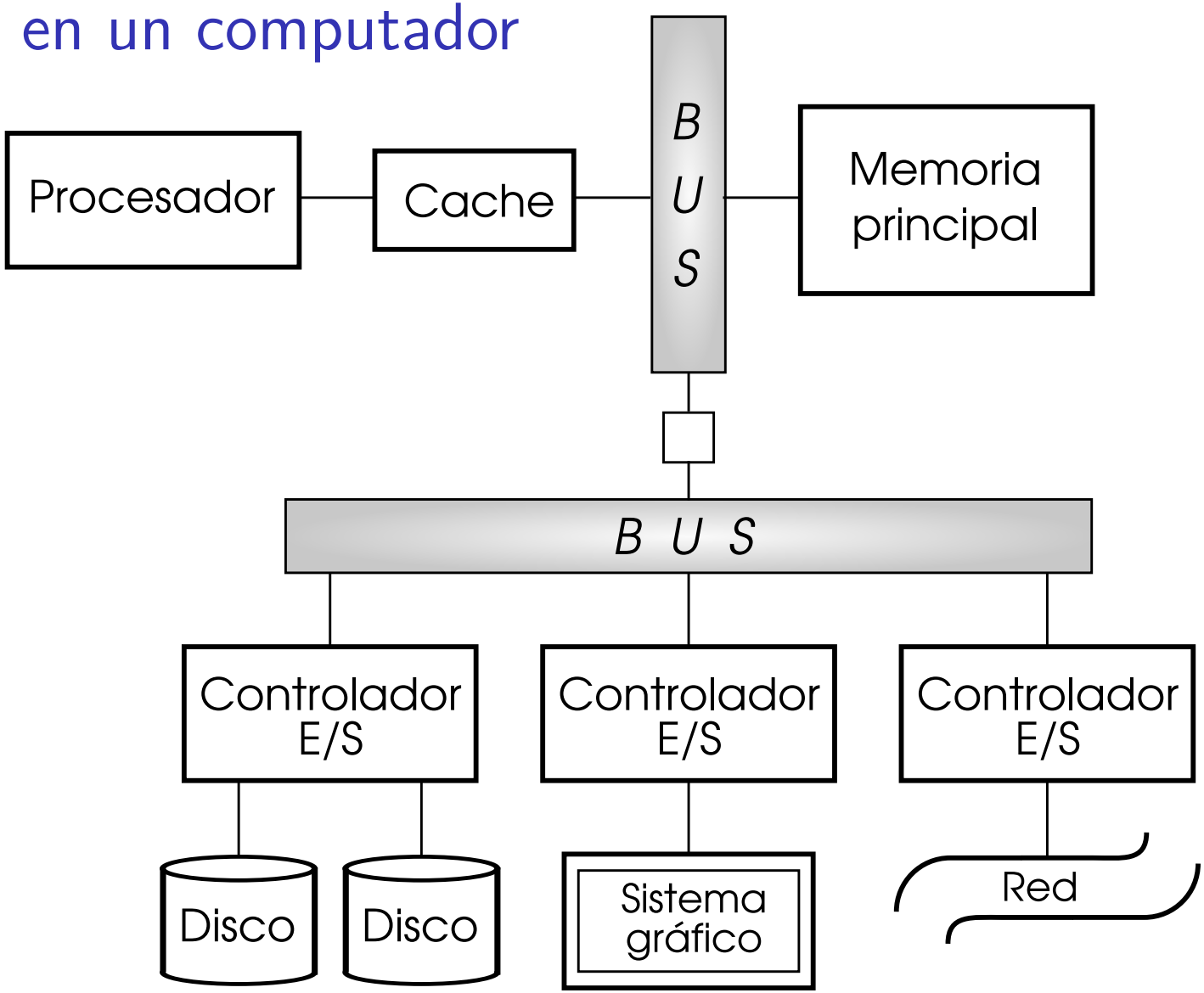
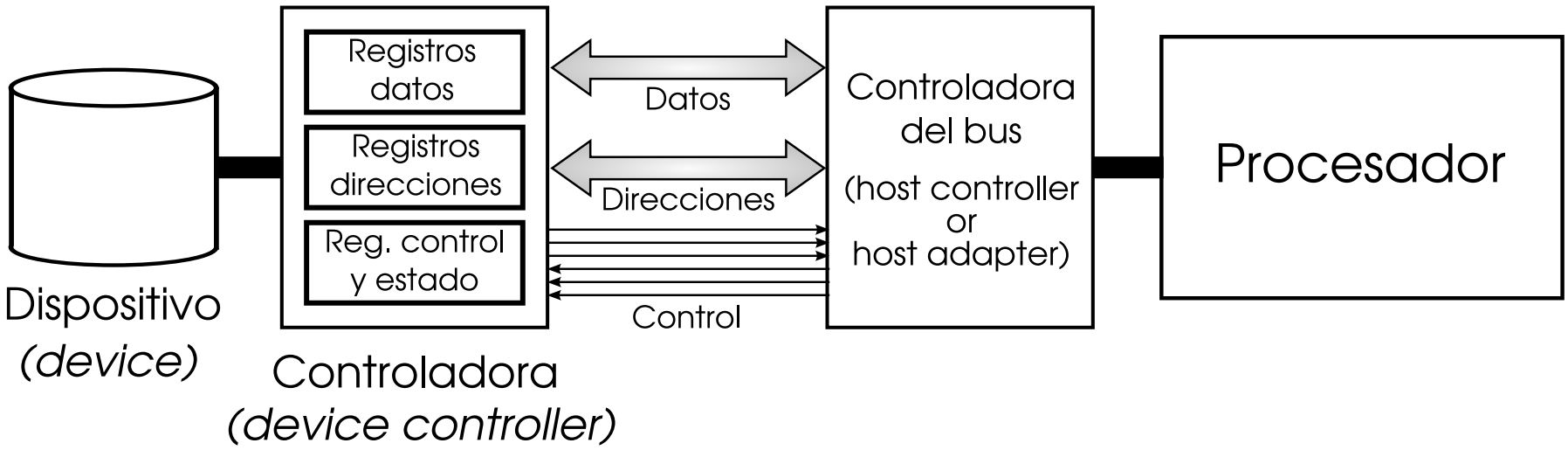


Figura: Componentes básicos de un computador: CPU, memoria, dispositivos de E/S y sistema de interconexión

Características de la Entrada/Salida

- Respecto a procesador & memoria, mayor énfasis en
  - Fiabilidad (*dependability*) y tolerancia a fallos
  - Escalabilidad/expansibilidad (*expandability*) y diversidad
- Problemática:
  - Amplia variedad de dispositivos $\Rightarrow$ (muy) distintas formas de funcionamiento
  - Velocidad de transferencia de periféricos $<<$ velocidad CPU o memoria
  - Diferentes formatos y tamaños de datos
  - Aspectos de rendimiento más difíciles de analizar y medir

Dispositivo de E/S (*I/O dev*)



- Dos componentes/partes fundamentales:
  - Dispositivo físico
  - Controladora del dispositivo: interfaz entre dispositivo y el sistema. Funciones principales:
    - ★ control y temporización
    - ★ almacenamiento temporal de datos (*buffering*)
    - ★ detección de errores
- La comunicación con los dispositivos de E/S se realiza esencialmente mediante registros

Acceso y gestión E/S

Interfaz de E/S

¿Cómo direccionamos/accedemos a los registros de un dispositivo de E/S?

- E/S asignada en memoria (*Memory-mapped I/O*, *MMIO*)
- E/S aislada (*Isolated I/O*, *Instruction-based I/O*, *Port I/O* o *Port-mapped I/O*, *PMIO*)

Gestión de E/S

¿Cómo gestionamos las operaciones de E/S?

- E/S programada (por encuesta)
- E/S por interrupciones (IRQs)
- E/S con acceso directo a memoria (DMA)
- Procesadores de E/S

Acceso y gestión E/S

Interfaz de E/S

¿Cómo direccionamos/accedemos a los registros de un dispositivo de E/S?

- E/S asignada en memoria (*Memory-mapped I/O*, *MMIO*)
- E/S aislada (*Isolated I/O*, *Instruction-based I/O*, *Port I/O* o *Port-mapped I/O*, *PMIO*)

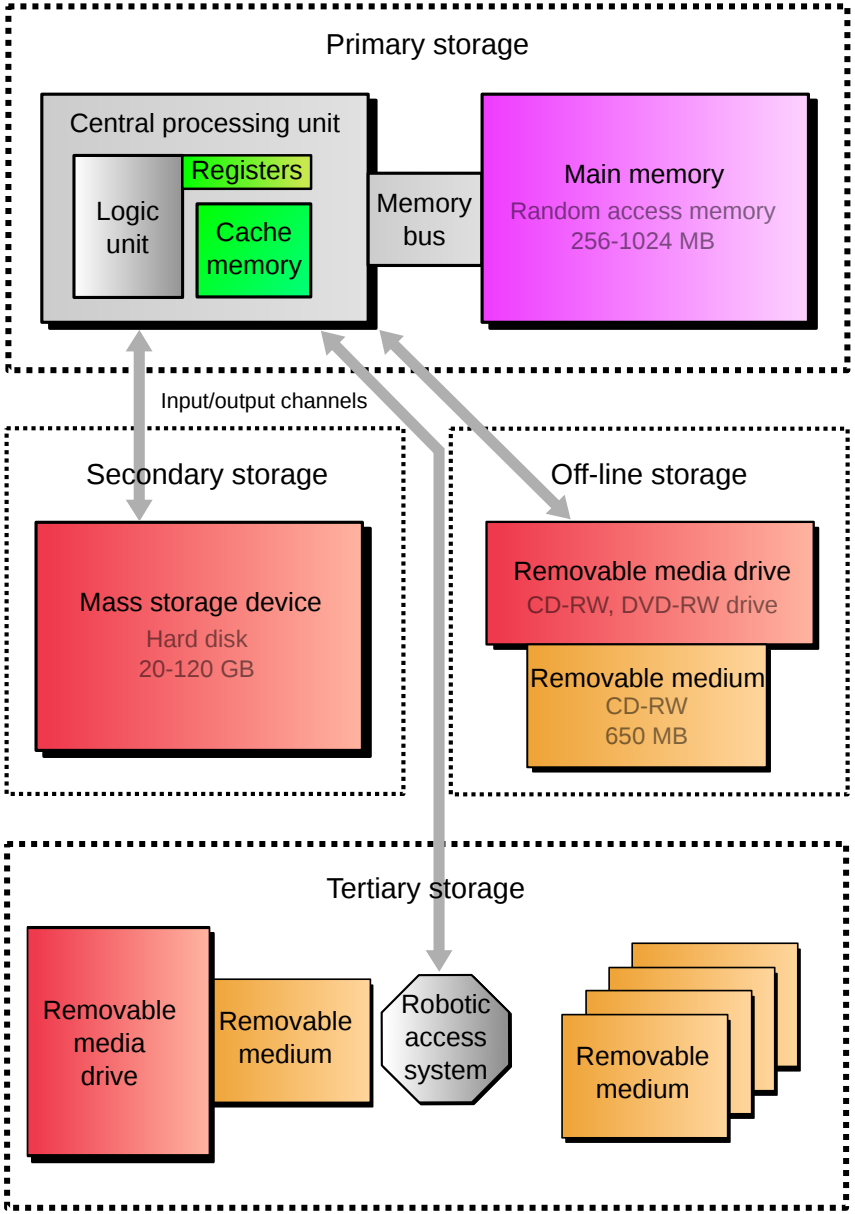
Gestión de E/S

¿Cómo gestionamos las operaciones de E/S?

- E/S programada (por encuesta)
- E/S por interrupciones (IRQs)
- E/S con acceso directo a memoria (DMA)
- Procesadores de E/S

Medios de almacenamiento

Jerarquía de memoria/almacenamiento

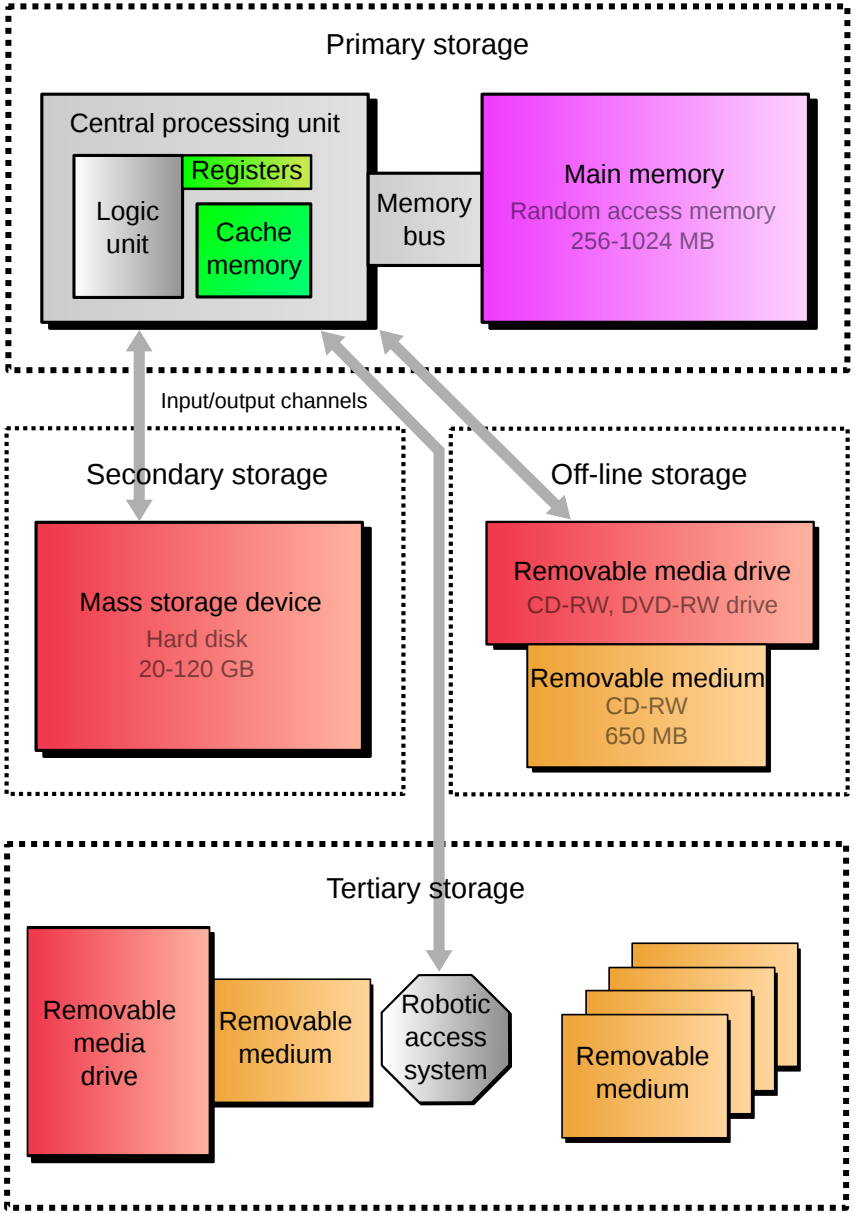


Distancia a CPU

- Más próximos
  - Más velocidad
  - Más coste
  - Menos capacidad
- Más alejados
  - Menos velocidad
  - Menos coste
  - Más capacidad

Medios de almacenamiento

Jerarquía de memoria/almacenamiento



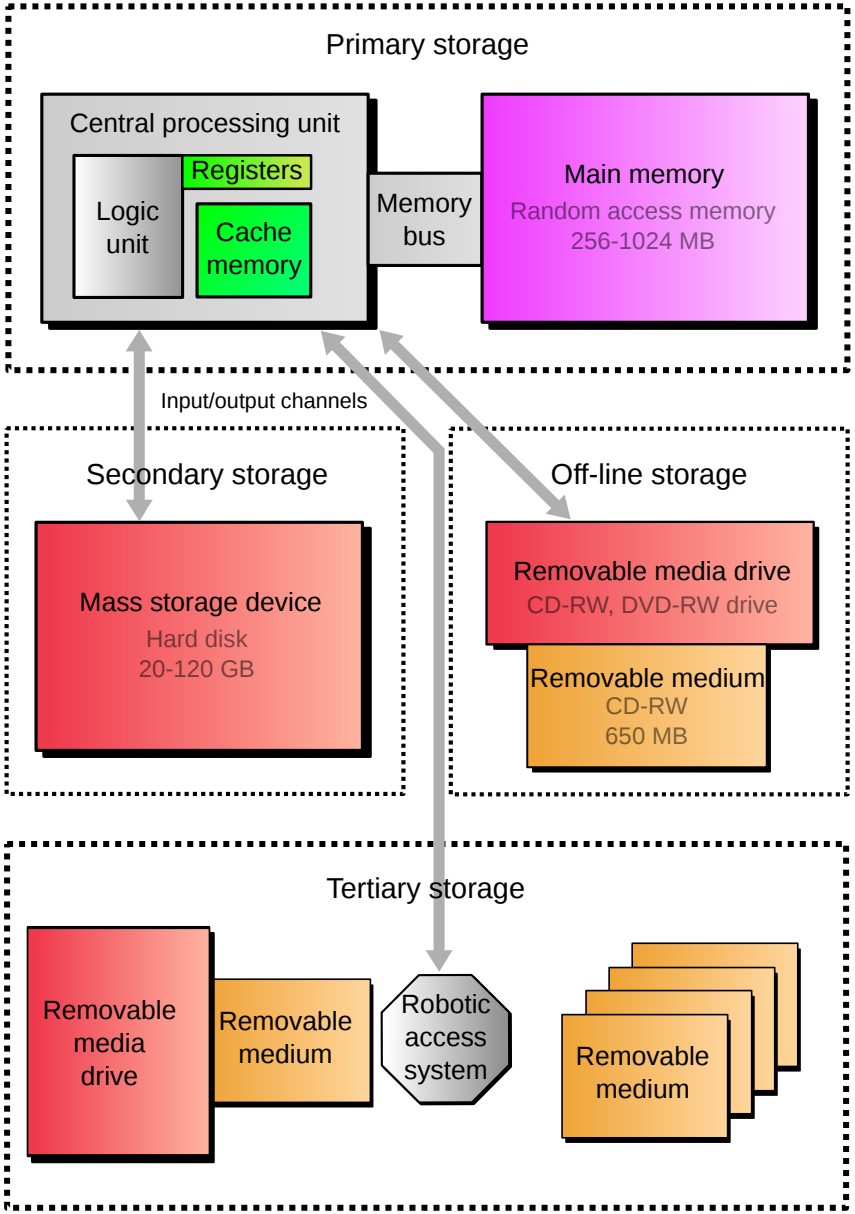
Distancia a CPU

- Más próximos
  - Más velocidad
  - Más coste
  - Menos capacidad
- Más alejados
  - Menos velocidad
  - Menos coste
  - Más capacidad

Tecno	Latencia	$/GB 2008
SRAM	0,5 – 2,5 ns	$2000 – $5000
DRAM	50 – 70 ns	$20 – $75
HDD	5 – 20 ms	$0,20 – $2

Medios de almacenamiento

Jerarquía de memoria/almacenamiento



Distancia a CPU

- Más próximos
  - Más velocidad
  - Más coste
  - Menos capacidad
- Más alejados
  - Menos velocidad
  - Menos coste
  - Más capacidad

Persistencia de información

- Volátiles
- No volátiles

Índice

Mass storage devices

- Dispositivos de E/S que componen memoria secundaria
  - Almacenamiento secundario, terciario y off-line
  - **Tradicionalmente** dispositivos de almacenamiento magnético
  - Respecto a almacenamiento primario (Regs, RAM, Cache)
    - ★ No volátiles
    - ★ Coste/Byte mucho menor
    - ★ Varios órdenes de magnitud más lentos
  - Base tecnologica más habitual
    - Magnética Disco duro (HDD), floppy, cinta...
    - Óptica CD, DVD, BD...
    - Semiconductores Memorias flash: SSD, pendrive

Mass storage devices

- Dispositivos de E/S que componen memoria secundaria
  - Almacenamiento secundario, terciario y off-line
  - **Tradicionalmente** dispositivos de almacenamiento magnético
  - Respecto a almacenamiento primario (Regs, RAM, Cache)
    - ★ No volátiles
    - ★ Coste/Byte mucho menor
    - ★ Varios órdenes de magnitud más lentos
  - Base tecnologica más habitual

    Magnética Disco duro (HDD), floppy, cinta...

    Óptica CD, DVD, BD...

    Semiconductores Memorias flash: SSD, pendrive

Mass storage devices

- Dispositivos de E/S que componen memoria secundaria
  - Almacenamiento secundario, terciario y off-line
  - **Tradicionalmente** dispositivos de almacenamiento magnético
  - Respecto a almacenamiento primario (Regs, RAM, Cache)
    - ★ No volátiles
    - ★ Coste/Byte mucho menor
    - ★ Varios órdenes de magnitud más lentos
  - Base tecnologica más habitual

    Magnética Disco duro (HDD), floppy, cinta...

    Óptica CD, DVD, BD...

    Semiconductores Memorias flash: SSD, pendrive

Mass storage devices

- Dispositivos de E/S que componen memoria secundaria
  - Almacenamiento secundario, terciario y off-line
  - **Tradicionalmente** dispositivos de almacenamiento magnético
  - Respecto a almacenamiento primario (Regs, RAM, Cache)
    - ★ No volátiles
    - ★ Coste/Byte mucho menor
    - ★ Varios órdenes de magnitud más lentos
  - Base tecnologica más habitual

    Magnética Disco duro (HDD), floppy, cinta...

    Óptica CD, DVD, BD...

    Semiconductores Memorias flash: SSD, pendrive

Almacenamiento magnético

Base tecnológica

Superficie magnetizable que con su interacción con un cabezal permite almacenar/obtener datos, haciendo las veces de memoria no volátil

- Principales dispositivos de almacenamiento magnético:

  Cinta (*Tape Drive*)
  - Acceso secuencial a los datos
  - Usadas para almacenamiento *offline* y *backup*

  Disco duro (*Hard Disk Drive, HDD*)
  - Base de la memoria secundaria
  - Perdiendo terreno ante tecnologías basadas en mem. flash

Presente-futuro: reemplazo de HDD con SSD

- Discos de estado sólido (*solid-state drives, SSD*)
  - Totalmente basados en componentes electrónicos, no tienen componentes mecánicos
  - Normalmente basados en memoria flash no volátil
    - También soluciones basadas en DRAM (volátil, necesaria batería)
  - Ventajas
    - Muy rápidos
    - Muy silenciosos
    - Más robustos (no fallos mecánicos)
    - Comportamiento casi determinista (tiempo de acceso constante)
    - Menor consumo eléctrico[*nosiempre*]
    - Pequeños y ligeros[*nosiempre*]
  - Desventajas
    - Precio
    - Capacidad
    - Más vulnerables a campos eléctricos y otros efectos
    - Vida más limitada[*mejorando*]

Índice

RAID *Redundant Array of Inexpensive/Independent Disks*)

Conjuntos de discos que operan independientemente y en paralelo, pero son vistos por el SO como un único dispositivo

Mejora en

- rendimiento
  - ✓ Con varios discos, **varias peticiones diferentes de E/S** pueden gestionarse en paralelo si datos requeridos residen físicamente en discos diferentes
  - ✓ Una **única petición de E/S** también puede tener acceso paralelo si el bloque de datos está distribuido a lo largo de varios discos (*stripping*)
  - ✗ En principio, con un conjunto de discos **fiabilidad disminuye**
    - ⋆ N discos tendrían $1/N$ veces la fiabilidad de un único disco
- fiabilidad
  - ✓ La fiabilidad pude incrementarse añadiendo información redundante: tolerancia a fallos.
  - ✓ Con redundancia, la fiabilidad de un conjunto de discos puede ser mucho mayor que la de un único disco grande equivalente
    - ⋆ MTTR (*mean time to repair*) << MTTF (*mean time to failure*)
  - ▶ Consejo práctico (basado en estudios experimentales): es buena idea no utilizar discos de la misma serie de fabricación en un mismo RAID

RAID *Redundant Array of Inexpensive/Independent Disks*)

Conjuntos de discos que operan independientemente y en paralelo, pero son vistos por el SO como un único dispositivo

Mejora en

- rendimiento
  - ✓ Con varios discos, **varias peticiones diferentes de E/S** pueden gestionarse en paralelo si datos requeridos residen físicamente en discos diferentes
  - ✓ Una **única petición de E/S** también puede tener acceso paralelo si el bloque de datos está distribuido a lo largo de varios discos (*stripping*)
  - ✗ En principio, con un conjunto de discos **fiabilidad disminuye**
    - ⋆ N discos tendrían $1/N$ veces la fiabilidad de un único disco
- fiabilidad
  - ✓ La fiabilidad pude incrementarse añadiendo información redundante: tolerancia a fallos.
  - ✓ Con redundancia, la fiabilidad de un conjunto de discos puede ser mucho mayor que la de un único disco grande equivalente
    - ⋆ MTTR (*mean time to repair*) $<<$ MTTF (*mean time to failure*)
  - ▶ Consejo práctico (basado en estudios experimentales): es buena idea no utilizar discos de la misma serie de fabricación en un mismo RAID

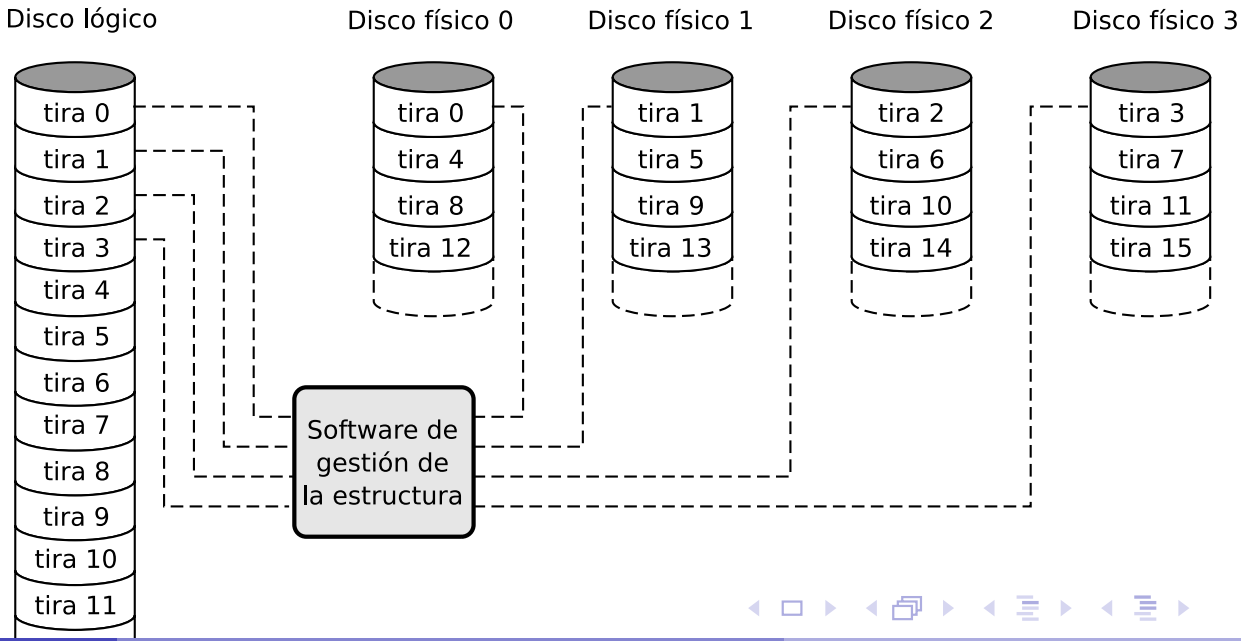
RAID: Características principales

- Variedad de alternativas para organizar los datos en múltiples discos
  - Toda una serie de esquemas estándares, con distintos grados de redundancia y rendimiento: **RAID**
- RAID: conjunto de esquemas o niveles independientes con las siguientes características comunes:
  1. Conjunto de unidades físicas de disco vistas por SO como una única unidad lógica
  2. Datos distribuidos a través de las unidades físicas del conjunto
  3. Redundancia de datos aumenta fiabilidad del conjunto
     Consideraciones de diseño (para reducir MTTR):
     - ★ discos de reserva (*hot spares*)
     - ★ cambio de discos en caliente (*hot swapping*)

RAID: principales esquemas/niveles

RAID 0

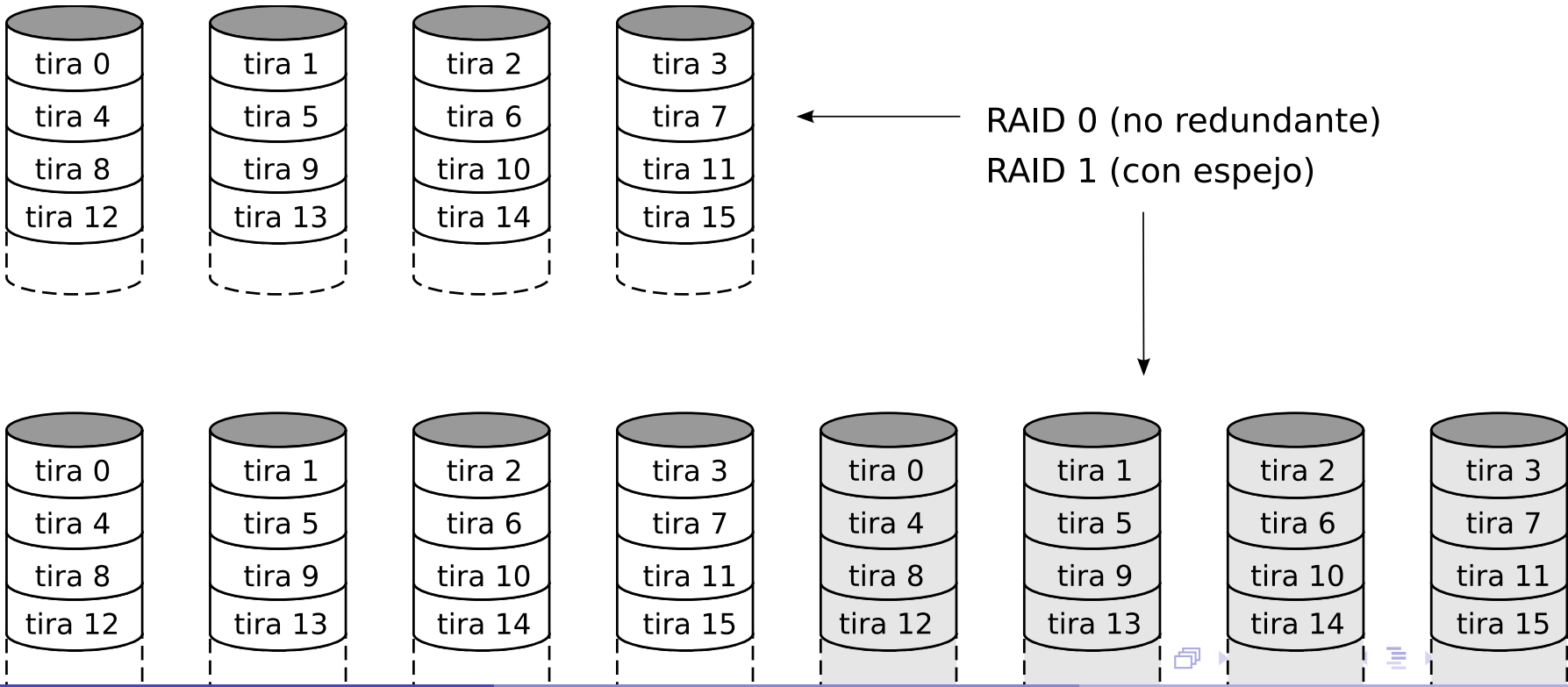
- No incluye redundancia de datos
- + prestaciones y capacidad a bajo coste, - fiabilidad
- Stripping: Información troceada en **tiras de datos** repartidas cíclicamente entre discos
- Franja: Conjunto de tiras lógicamente consecutivas proyectadas sobre misma tira en cada disco
- Op. E/S que implica a tiras lógicas contiguas (misma franja): acceso paralelo a discos $\Rightarrow$ reducción T_{transf}
- Array de discos presentado como 1 único disco grande



RAID: principales esquemas/niveles (y II)

RAID 1

- Todos esquemas RAID salvo 0 incluyen información redundante para permitir cierta recuperación de datos
- En RAID 1 redundancia se logra con duplicación de todos los datos
- Distribución cíclica de datos, como en RAID 0, pero con un disco espejo para cada disco del conjunto
- Gran fiabilidad



RAID: principales esquemas/niveles (y III)

RAID 1

Ventajas:

- ✓ Una petición de lectura puede ser servida por cualquiera de los discos que contiene los datos pedidos
- ✓ Respecto a RAID 0: posibilidad de recuperar errores
- ✓ Respecto a RAID 2-5: apenas hay penalización de escritura
- ✓ La recuperación tras un fallo es sencilla

Desventajas:

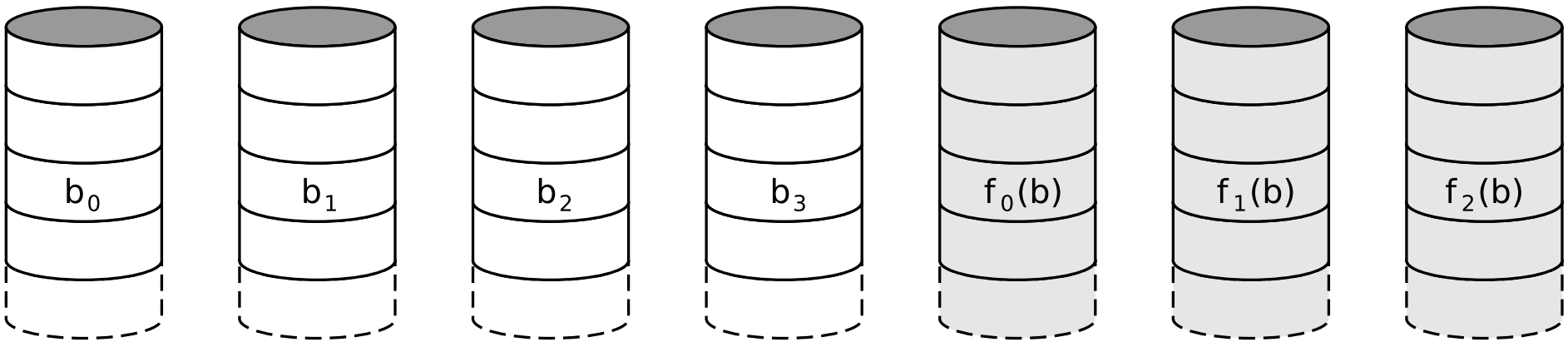
- ✗ Coste almacenamiento: requiere el doble del espacio del disco lógico que se quiere soportar

RAID: principales esquemas/niveles (y IV)

- Demás niveles RAID no duplican toda la información
  - Códigos de detección y corrección de errores (*ECC*)
  - Menos información redundante $\Rightarrow$ Menos coste almacenamiento

RAID 2

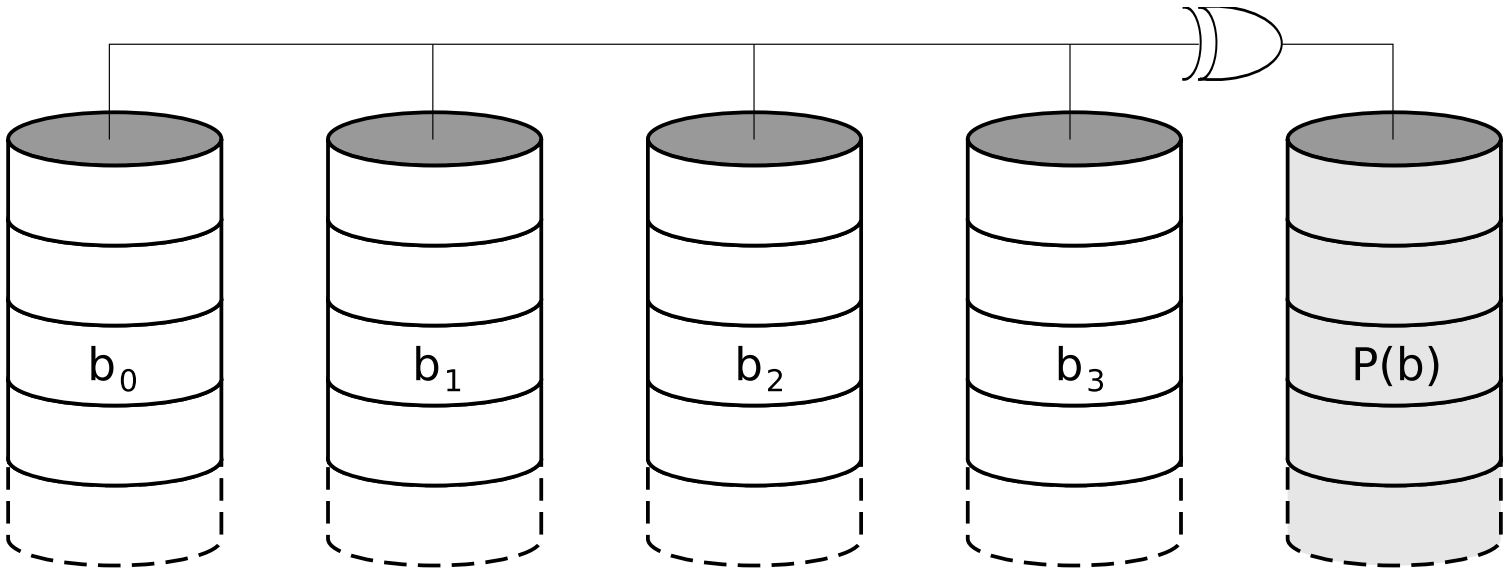
- ECC: Códigos de Hamming, que se almacenan en discos de redundancia a parte de los datos
- Stripping de datos en tiras muy pequeñas (a nivel de byte, incluso)
- Discos sincronizados, realizando misma operación E/S
- Altas velocidades de transferencia, aunque excesivamente complejo y costoso (no se implementa en la práctica)



RAID: principales esquemas/niveles (y V)

RAID 3

- También stripping de datos en tiras muy pequeñas (a nivel de byte, incluso)
- También discos sincronizados, realizando en paralelo la misma operación E/S
- Diferencia con RAID 2: solo un disco de redundancia, usando **información de paridad** como ECC
- Alta velocidad de transferencia en cada transacción (como RAID 2, solo una operación de E/S a la vez)
- Implementación bastante costosa



RAID: principales esquemas/niveles (y VI)

- Cálculo del bit de paridad i-ésimo:

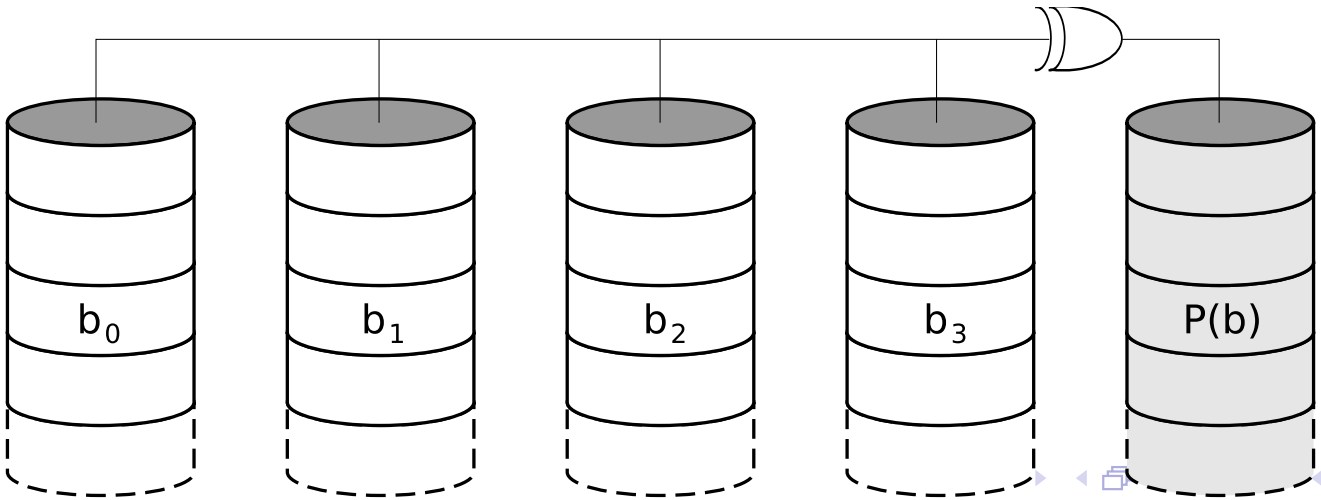
$$X4(i) = X3(i) \oplus X2(i) \oplus X1(i) \oplus X0(i)$$

Recuperación bit i-ésimo en caso de fallo del disco $X1$:

$$X1(i) = X4(i) \oplus X3(i) \oplus X2(i) \oplus X0(i)$$

Penalización de escritura: dos lecturas y dos escrituras

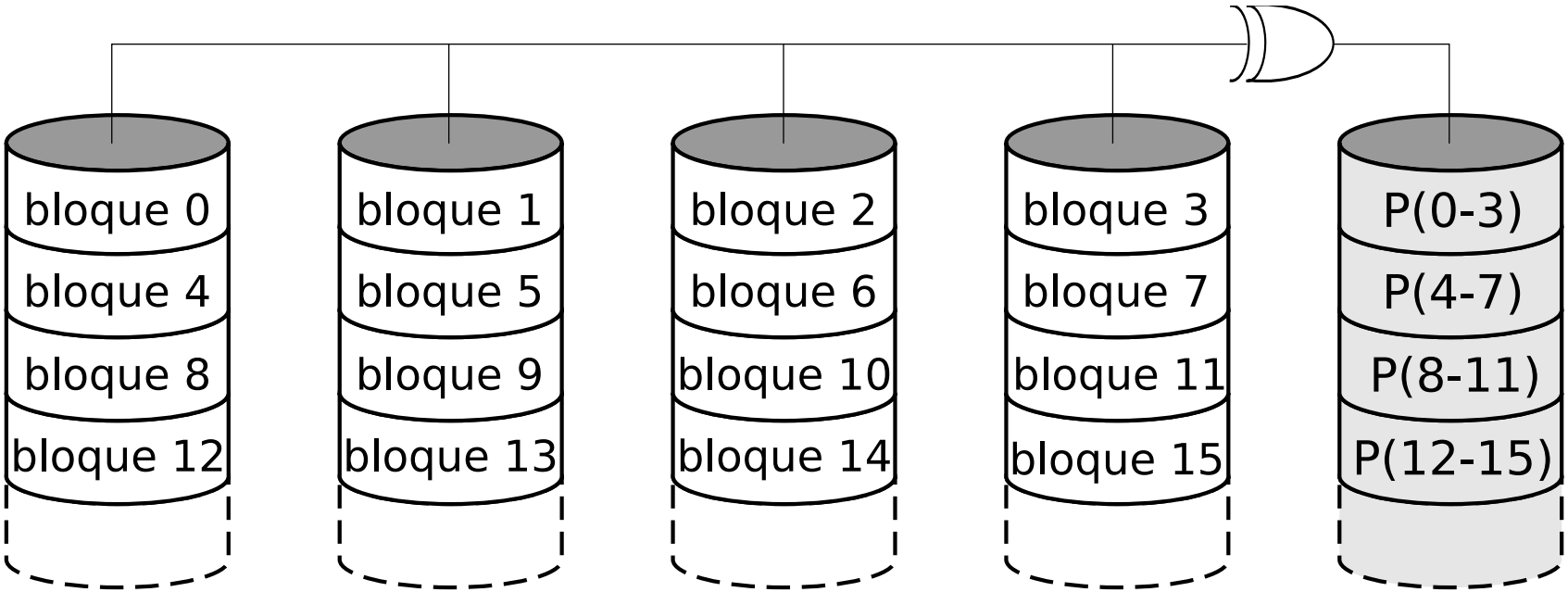
$$\begin{aligned} X4'(i) &= X3(i) \oplus X2(i) \oplus X1'(i) \oplus X0(i) \\ &= X3(i) \oplus X2(i) \oplus X1'(i) \oplus X0(i) \oplus X1(i) \oplus X1(i) \\ &= X4(i) \oplus X1(i) \oplus X1'(i) \end{aligned}$$



RAID: principales esquemas/niveles (y VII)

RAID 4

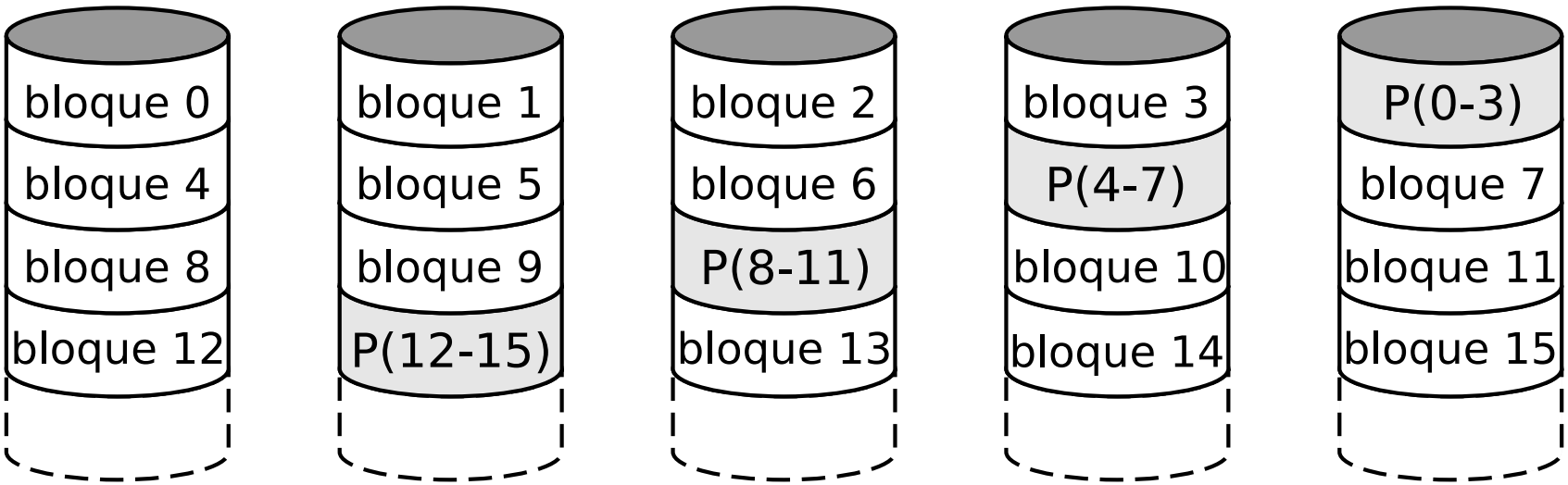
- Acceso independiente a los discos
  - ▶ peticiones de E/S separadas se atienden en paralelo
- Tiras de datos de mayor tamaño que en RAID 2 y 3
- Tiras con bits de paridad en **un** disco de paridad
  - ✗ cuello de botella en acceso a ese disco



RAID: principales esquemas/niveles (y VIII)

RAID 5

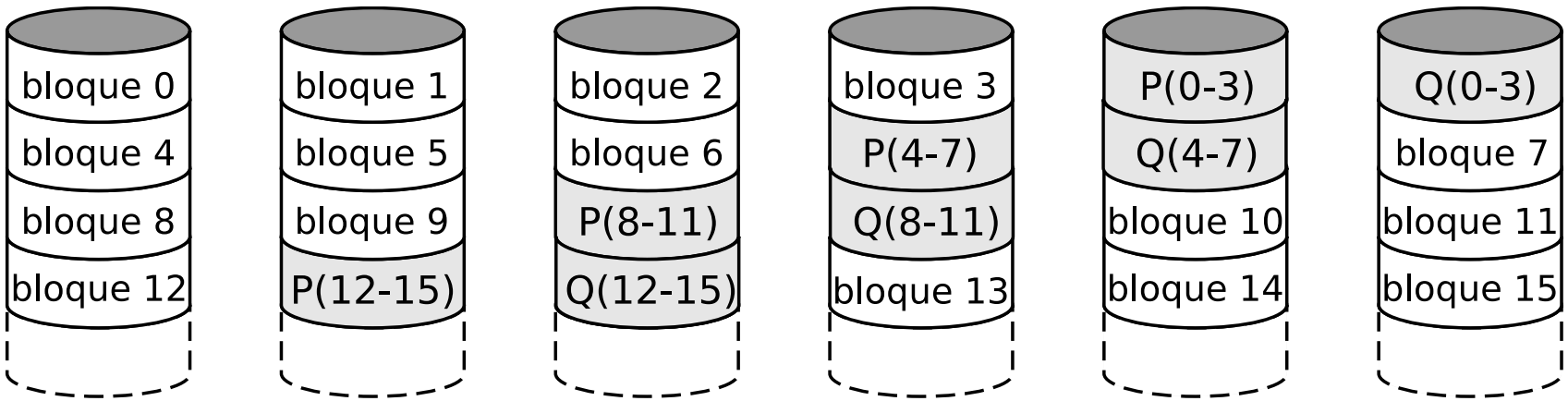
- Como RAID 4, pero distribuyendo tiras de paridad a lo largo de todos los discos (cíclicamente)
- Eliminamos cuello de botella de RAID 4



RAID: principales esquemas/niveles (y IX)

RAID 6

- Como RAID 5, pero con dos *ECCs* en discos diferentes
  - *paridad* + *paridad* o *paridad* + *otroECC*
- Permite recuperarse de dos fallos de disco simultáneos



RAID: resumen esquemas/niveles

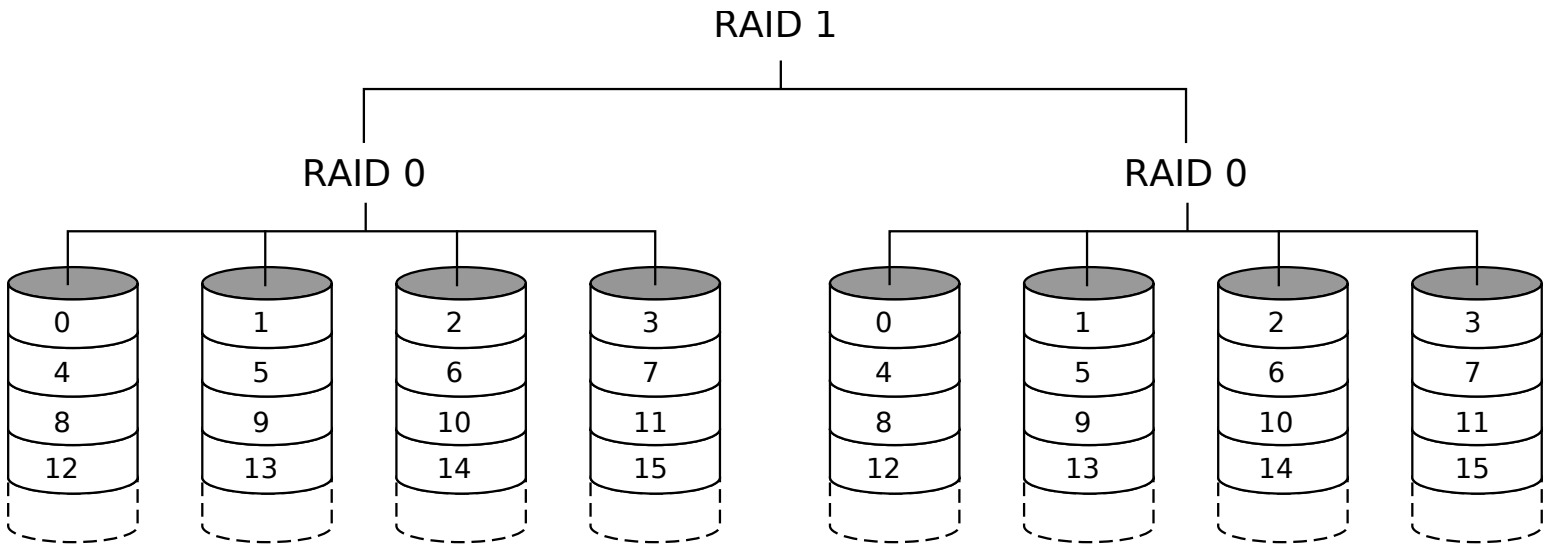
- Respecto a fiabilidad/información redundante
  - RAID 0 no tiene redundancia, **no** es realmente *«RAID»*
  - RAID 1 duplica toda la información (espejo)
  - Resto de niveles RAID incrementan fiabilidad mediante esquemas de detección de errores, sin necesidad de duplicar toda la información
    - ★ Menor cantidad de información redundante que en RAID 1
    - ★ Menor fiabilidad que RAID 1, pero mejor relación seguridad/precio
  - En RAID 2, 3 y 4 información redundante en un único disco
  - En RAID 5 y 6 info redundante distribuida entre todos los discos
- Respecto a rendimiento
  - RAID 2 y 3: una única op. E/S que accede en paralelo a todos los discos
    - ★ Tiras muy pequeñas
    - ★ Incrementamos velocidad de transferencia (disminuimos latencia) de cada operación
  - RAID 4, 5 y 6: varias ops. E/S concurrentes
    - ★ Tiras de mayor tamaño
    - ★ Aumentamos *nžops/s*

RAID: resumen esquemas/niveles

- Respecto a fiabilidad/información redundante
  - RAID 0 no tiene redundancia, **no** es realmente *«RAID»*
  - RAID 1 duplica toda la información (espejo)
  - Resto de niveles RAID incrementan fiabilidad mediante esquemas de detección de errores, sin necesidad de duplicar toda la información
    - ★ Menor cantidad de información redundante que en RAID 1
    - ★ Menor fiabilidad que RAID 1, pero mejor relación seguridad/precio
  - En RAID 2, 3 y 4 información redundante en un único disco
  - En RAID 5 y 6 info redundante distribuida entre todos los discos
- Respecto a rendimiento
  - RAID 2 y 3: una única op. E/S que accede en paralelo a todos los discos
    - ★ Tiras muy pequeñas
    - ★ Incrementamos velocidad de transferencia (disminuimos latencia) de cada operación
  - RAID 4, 5 y 6: varias ops. E/S concurrentes
    - ★ Tiras de mayor tamaño
    - ★ Aumentamos $nžops/s$

RAID: esquemas combinados/anidados

RAID 0+1



RAID 1+0