Tema II

Unidad de memoria

2 Unidad de memoria

2.1 Definiciones y conceptos básicos
- 2.1.1 Localización
- 2.1.2 Capacidad
- 2.1.3 Unidad de transferencia
- 2.1.4 Método de acceso
- 2.1.5 Tipos físicos
- 2.1.6 Características físicas
- 2.1.7 Velocidad
- 2.1.8 Organización
- 2.1.9 Resumen de características propiedades de la memoria

2.2 Jerarquía de memorias
- 2.2.1 Ejemplo: Sistema con dos niveles de memoria

2.3 Memorias de semiconductor
- 2.3.1 Características generales de un CIM
- 2.3.2 Ejemplo: Cálculo del número de ciclos de reloj en los accesos a memoria
- 2.3.3 Estructura de la celda básica de memoria
- 2.3.4 Organización interna
- 2.3.5 Diseño de bloques de memoria
- 2.3.6 Conexión de la unidad de memoria al bus del sistema
- 2.3.7 Estructura y direccionamiento de la unidad de memoria

2.4 Memorias caché
- 2.4.1 Rendimiento de una memoria caché
- 2.4.2 Capacidad de la memoria caché
- 2.4.3 Organización de la memoria caché
- 2.4.4 Algoritmos de reemplazamiento
- 2.4.5 Estrategia de escritura
- 2.4.6 Tamaño del bloque
- 2.4.7 Número de cachés

2.5 Memorias asociativas
- 2.5.1 Ejemplo: Concepto de memoria asociativa
- 2.5.2 Estructura de una memoria asociativa
- 2.5.3 Determinación de la función lógica del registro de máscara
- 2.5.4 Operación de lectura
- 2.5.5 Operación de escritura
- 2.5.6 Ejemplo: Diseño de una memoria asociativa

2.8 Discos magnéticos
- 2.8.1 Estructura física
- 2.8.2 Ejemplo: Tiempo de acceso a un archivo de acceso secuencial y aleatorio
- 2.8.3 Controlador del disco
- 2.8.4 Planificación del disco

NO SON OBJETO DE EXAMEN

2.6 Memorias compartidas (completo)

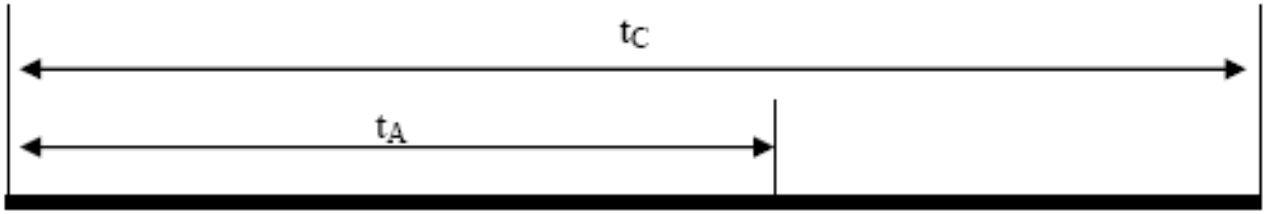
2.7 Memorias tipo pila (completo)

2.1 Definiciones y conceptos básicos

Localización	Memoria interna de la CPU Memoria principal Memoria secundaria	**Tipo físico**	Memoria de semiconductor Memoria magnética Memoria óptica
Capacidad	Número de palabras Tamaño de la palabra	**Características Físicas**	Alterabilidad (ROM, RAM) Permanencia de la información - DRO/NDRO - Volátil/No volátil - Estática/Dinámica
Unidad de transferencia	Palabra Bloque	**Velocidad**	Tiempo de acceso t_A Tiempo de ciclo t_C Frecuencia de acceso f_A
Método de acceso	Acceso aleatorio Acceso secuencial Acceso directo Acceso asociativo	**Organización**	2D 21/2D

Tabla 2.3: Características básicas de las memorias

Velocidad



Velocidad: Para medir el rendimiento se utilizan los siguientes parámetros:

a) Tiempo de acceso (T_A). Se define como el tiempo medio necesario para leer/escribir una cantidad fija de información.
b) Tiempo de ciclo de memoria (t_c) ; el intervalo de tiempo mínimo entre dos lecturas consecutivas, puede ser mayor que T_A .
c) Velocidad de transferencia o frecuencia de acceso (f_A): Es el número de palabras/segundo que pueden ser accedidas:

En el caso de acceso aleatorio : $f_A = 1/t_c$

En memorias de acceso no aleatorio: $t_n = t_A + n/p$

donde: t_n : Tiempo medio en leer/escribir n bits
t_A : Tiempo de acceso medio.
n : Número de bits.
p : Velocidad de transferencia (bits/seg.)

Características de los circuitos integrados de memorias

Característica	ROM	RAM estática	RAM dinámica
Alterable	No	Si	Si
Capacidad	Muy alta	Alta	Muy alta
Tiempo de acceso	Muy bajo	Muy bajo	Bajo
Refresco	No	No	Si
Volátil	No	Si	Si

Tabla 2.2: Características de los circuitos integrados de memoria

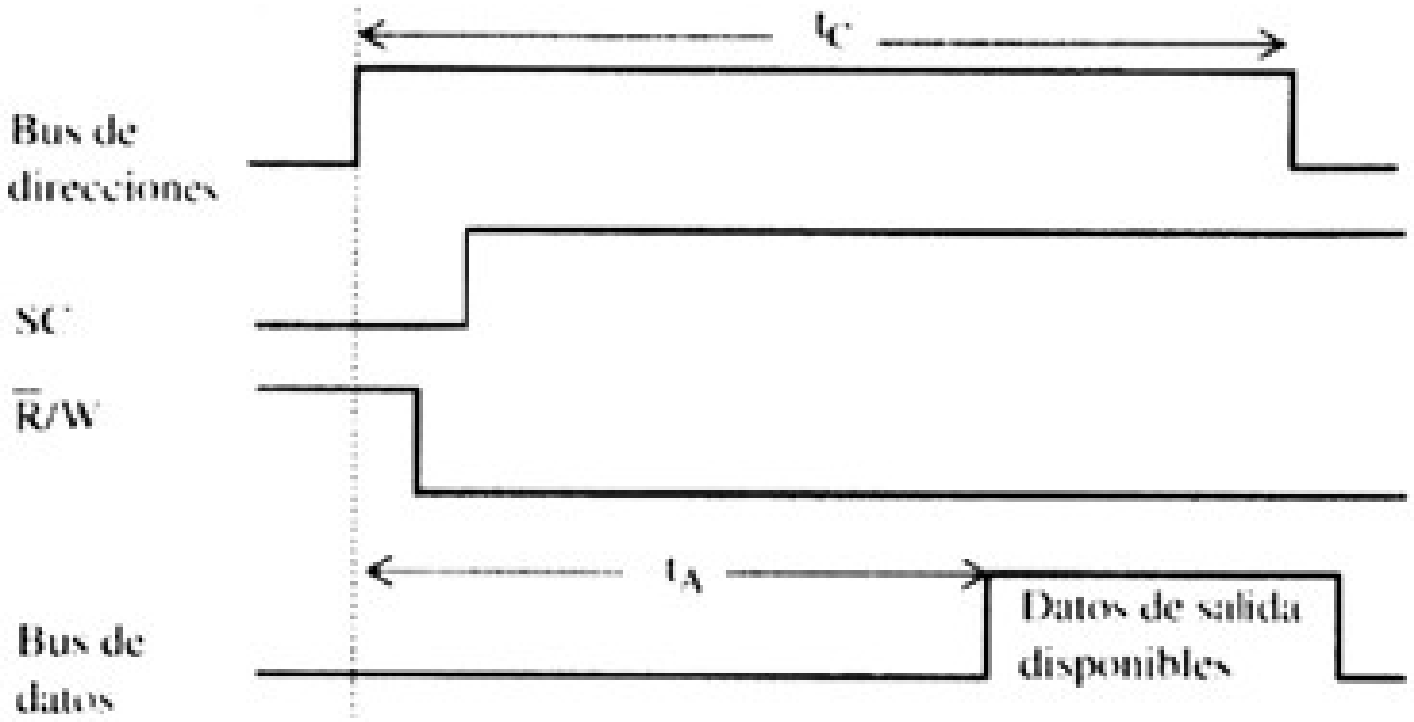


Figura 2.10: Ciclo de lectura de un CIM

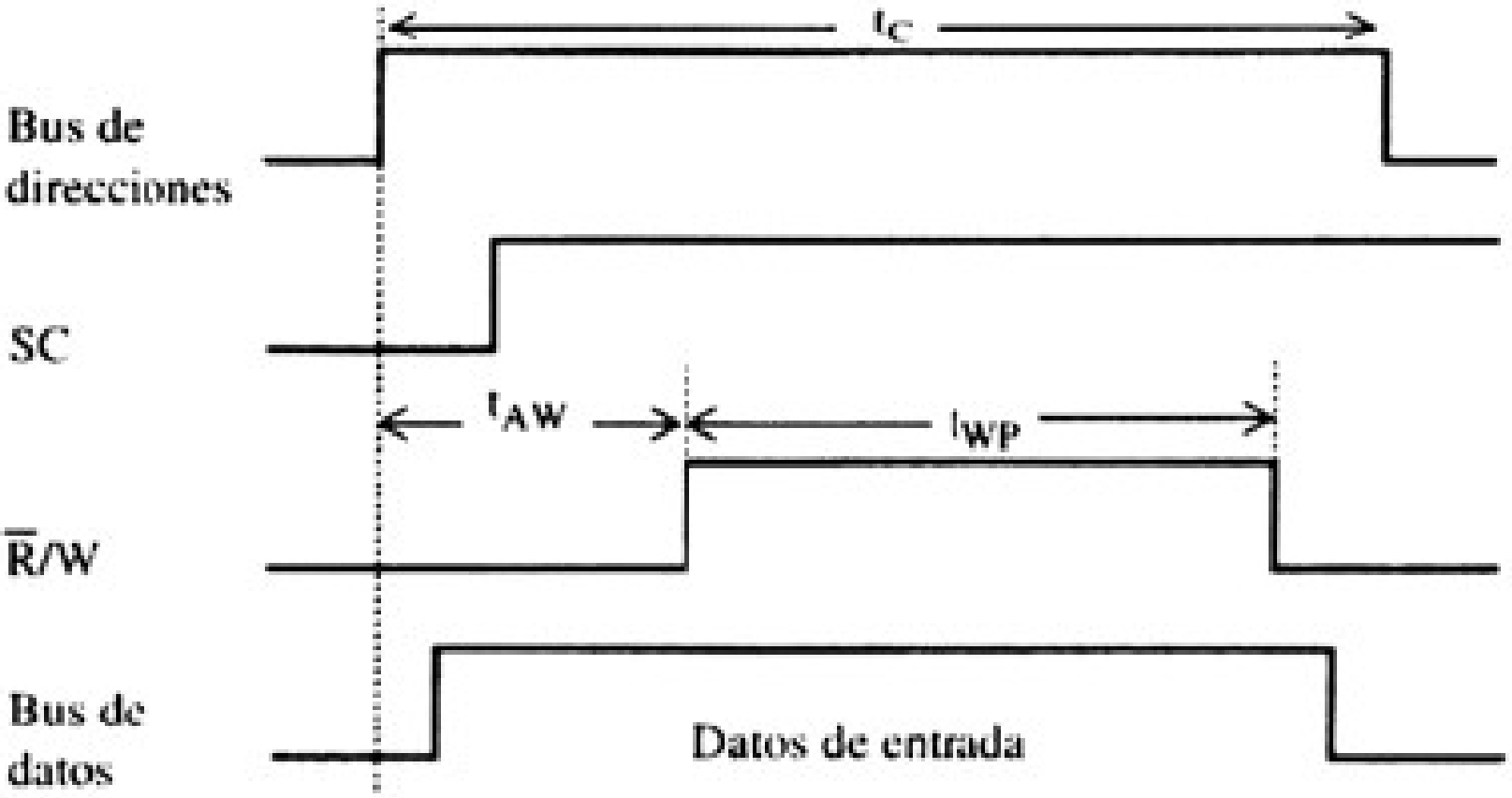
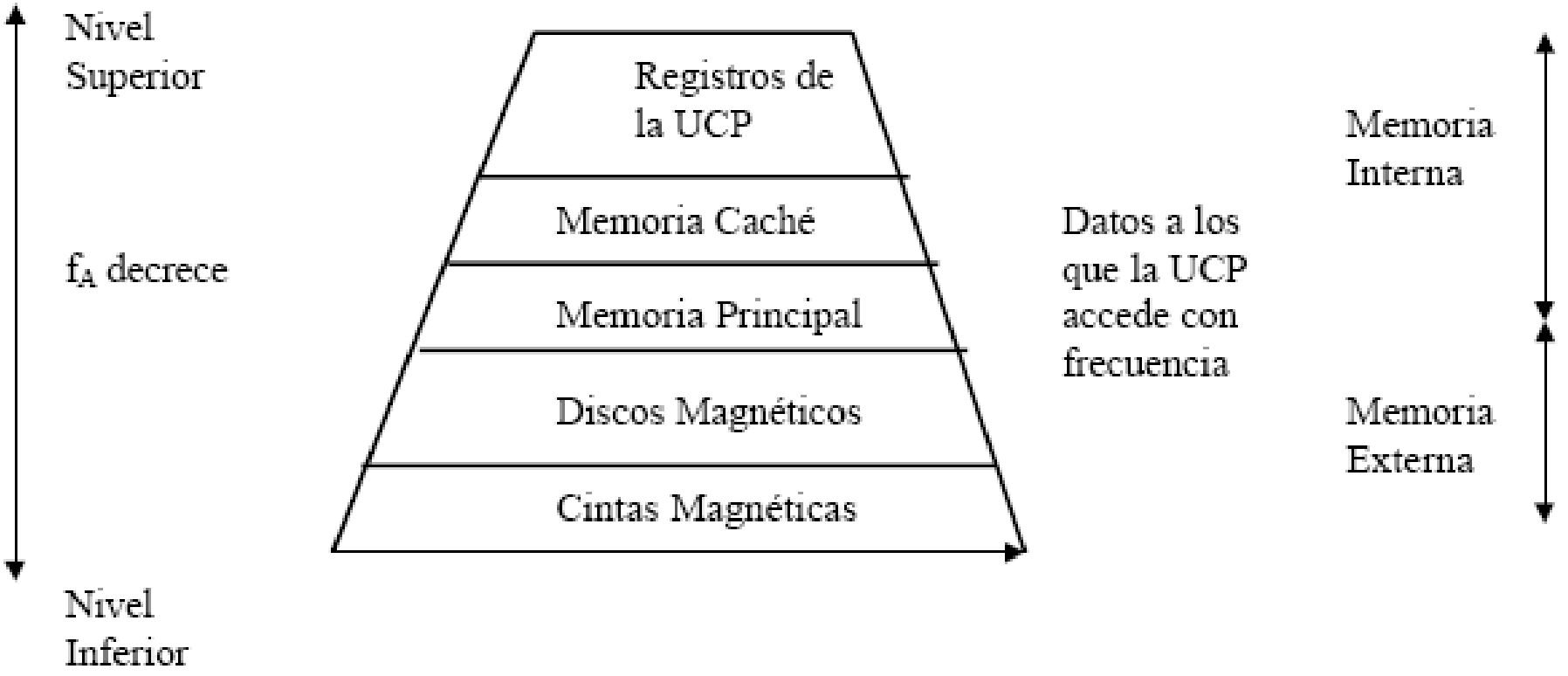


Figura 2.11: Ciclo de escritura de un CIM

2.2 Jerarquía de memorias



Parámetros en la jerarquía de las memorias

- Capacidad
- Velocidad
- Coste por bit
- Frecuencia de utilización
- **Principio de localidad referencia:**
  - Se entiende como el índice de probabilidad de uso de la información que está en memoria
  - Localidad temporal:
    - Tendencia a reutilizar los datos e instrucciones utilizados recientemente.
  - Localidad espacial:
    - Tendencia a referenciar las instrucciones y datos próximos a los que están utilizando

Principios

- Con la utilización de las jerarquías de memoria las memorias rápidas de baja capacidad y alto coste se complementan con las memorias lentas de gran capacidad y bajo coste.
- El tiempo de acceso medio total (tA) empleado por la UCP para acceder a una palabra se puede expresar por la relación:

-

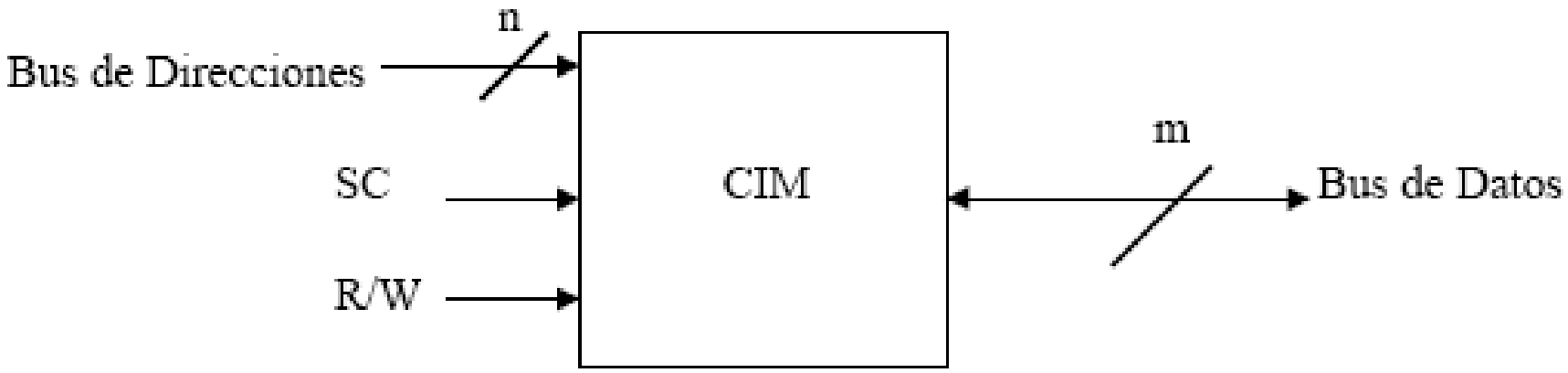
$$t_A = t_{A1} + t_{A2} - \frac{Tt_{A2}}{100}$$

donde:

t_{A1} es el tiempo de acceso a la memoria de nivel 1
t_{A2} es el tiempo de acceso a la memoria de nivel 2
T es el porcentaje de tiempo total en que la palabra a la que desea acceder la UCP se encuentra en la memoria de nivel 1.

2.3 Memorias de semiconductor

- Características
  - Un CIM está organizado internamente como una matriz de N x m celdas elementales, en la que se pueden almacenar N palabras de m bits. A cada palabra almacenada en el CIM se le asigna una única dirección.



Estructura de la celda básica de memoria

- El elemento básico de un CIM es la celda de memoria que permite almacenar un bit de información.
- Propiedades de las celdas de memoria de tipo semiconductor:
  - Presentan dos estados estables (o semiestables)
  - Se pueden escribir (al menos una vez) para fijar su estado
  - Se pueden leer para conocer su estado

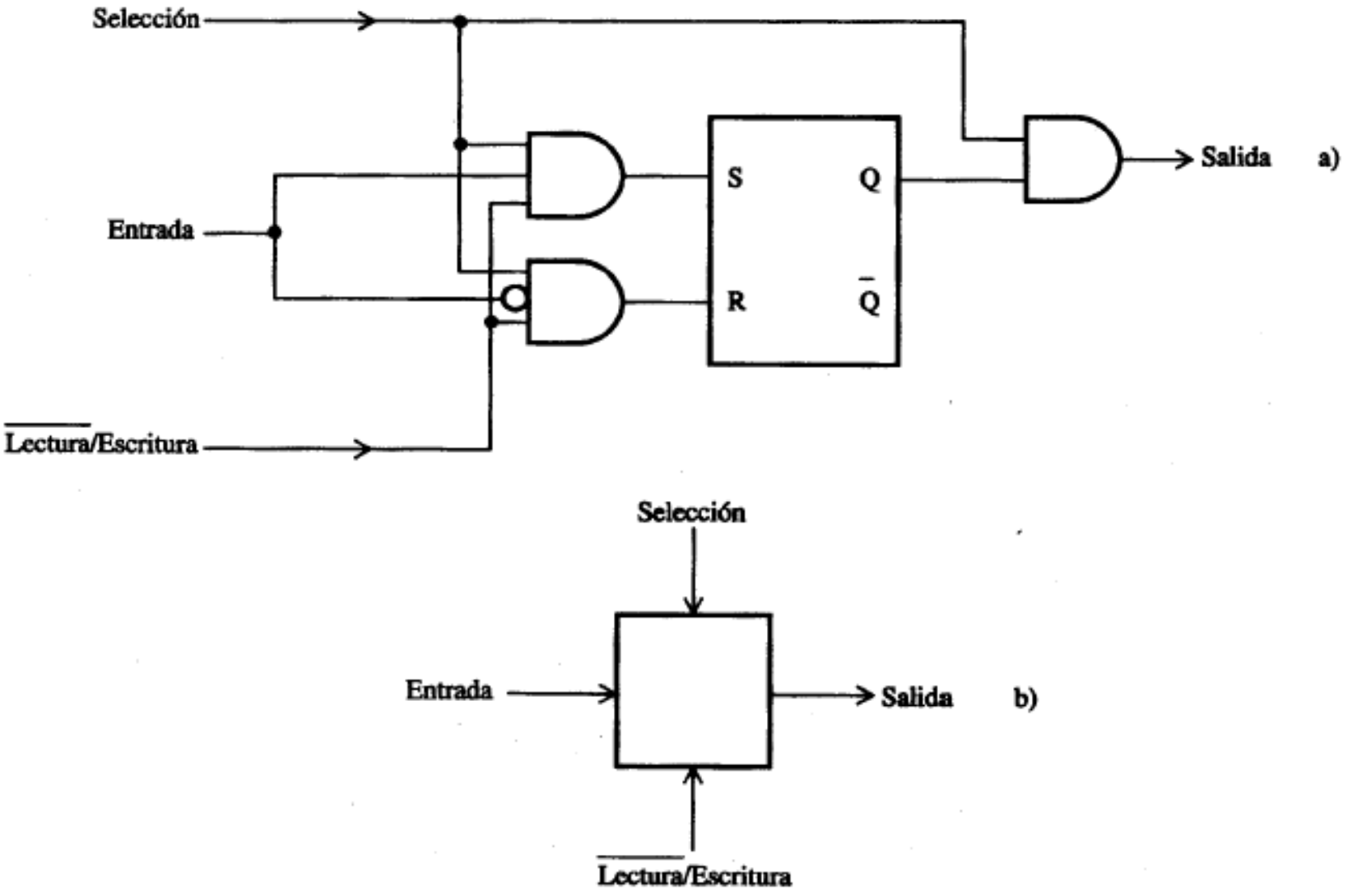


Figura 2.12: a) Celda básica de memoria RAM b) Esquema de la celda básica de memoria

Organización interna: 2D

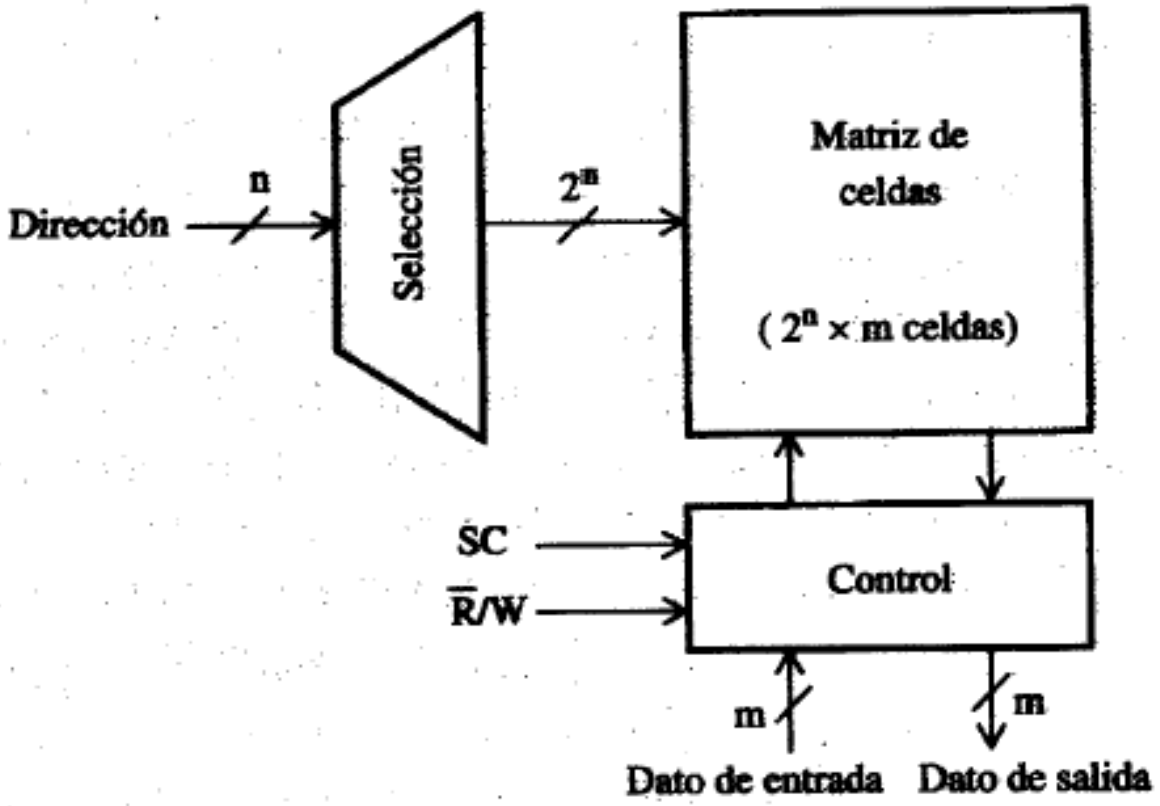


Figura 2.13: Organización 2D de una memoria RAM

La organización 2D da lugar a matrices de celdas excesivamente largas y estrechas que no son adecuadas para su realización en un circuito integrado.

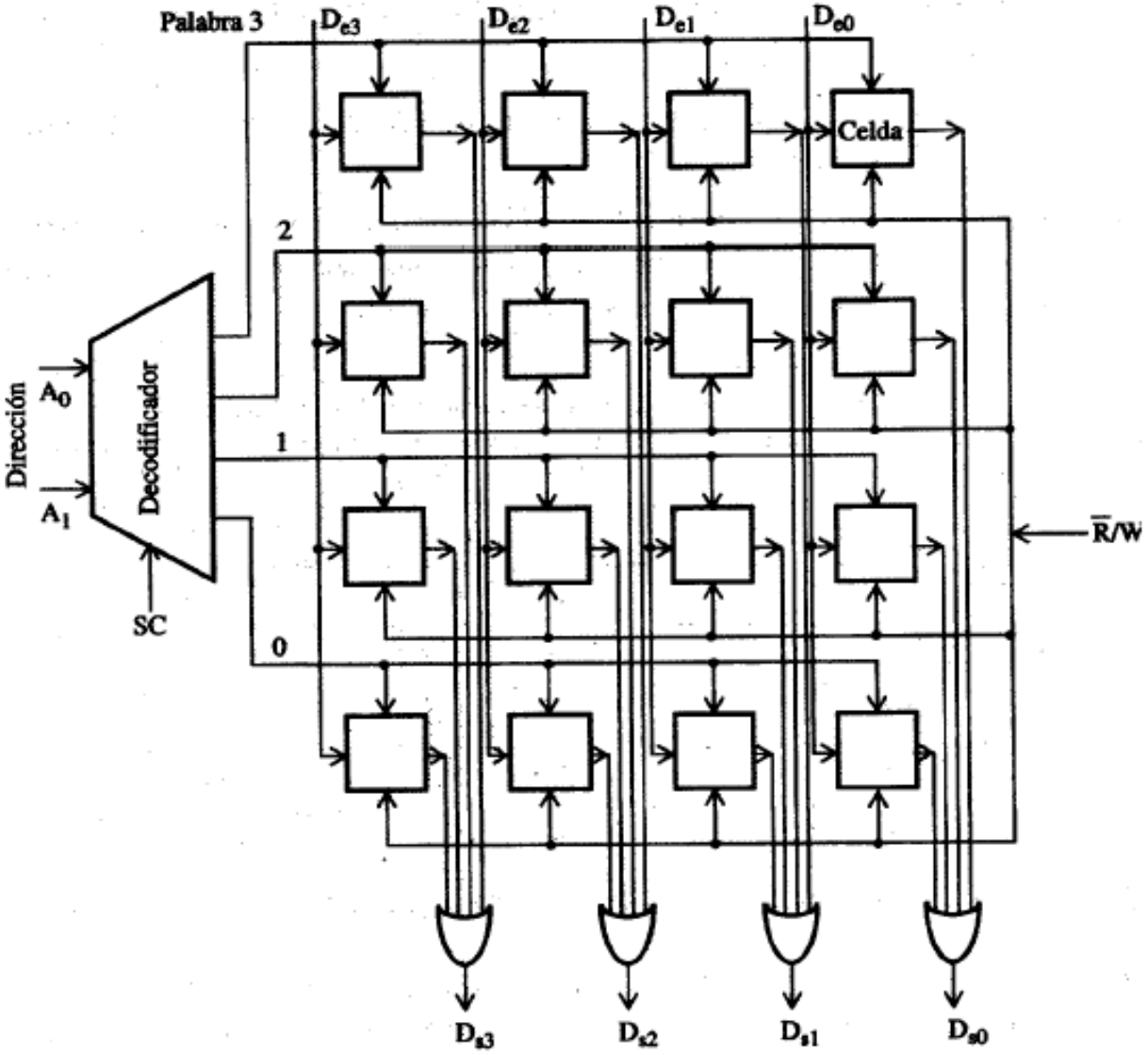


Figura 2.14: Memoria RAM de 4 palabras con 4 bits por palabra

Organización 2D 1/2

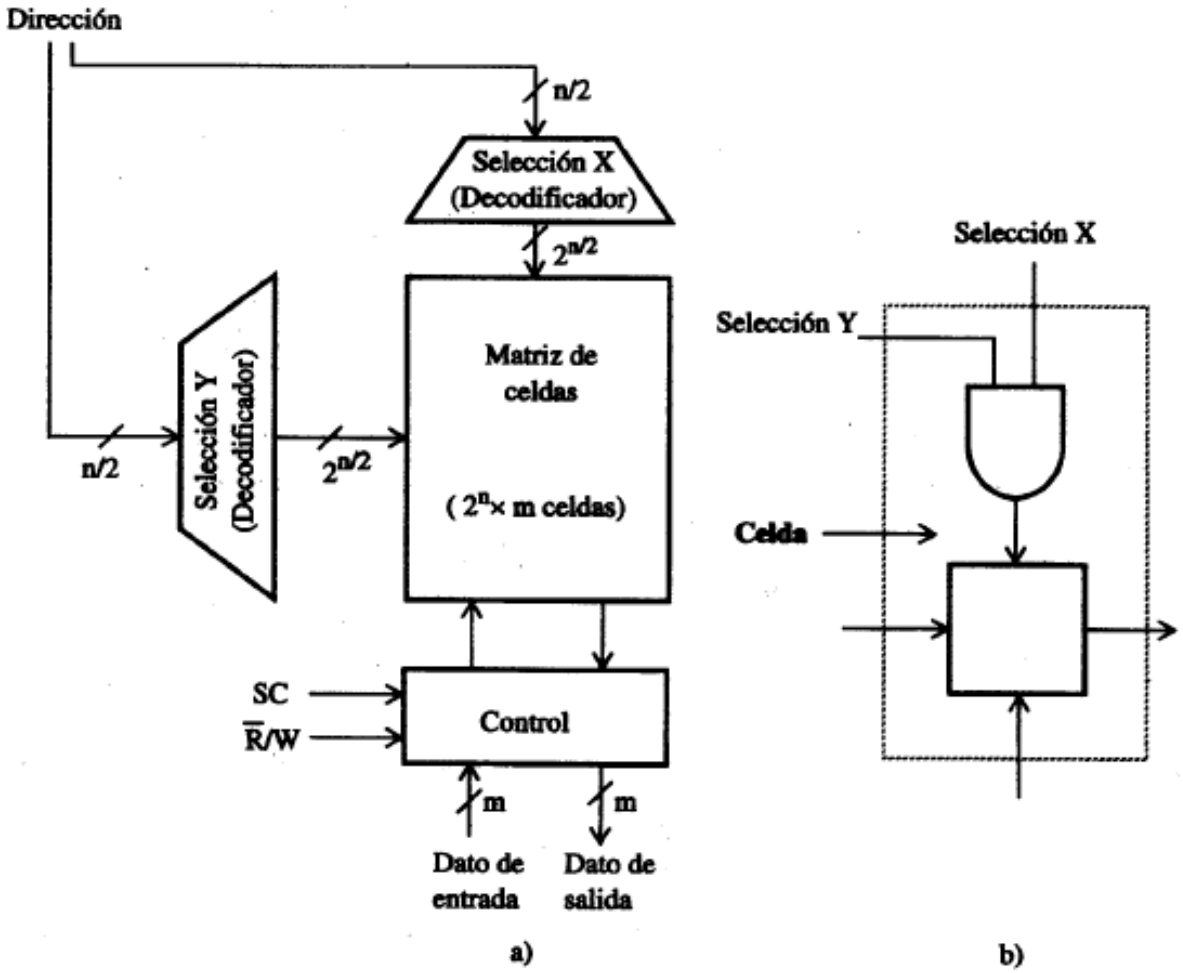


Figura 2.15: a) Memoria RAM con decodificación por coincidencia b) Celda básica de memoria modificada

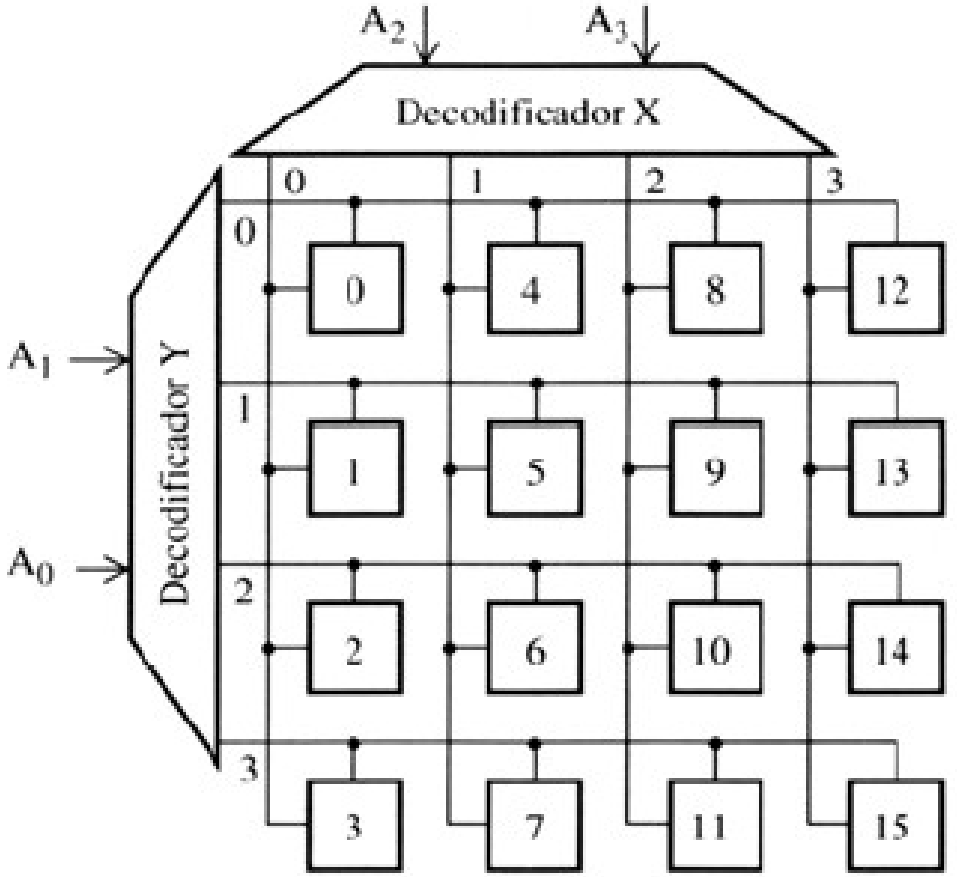


Figura 2.16: Memoria RAM de 16 × 1 con selección por coincidencia

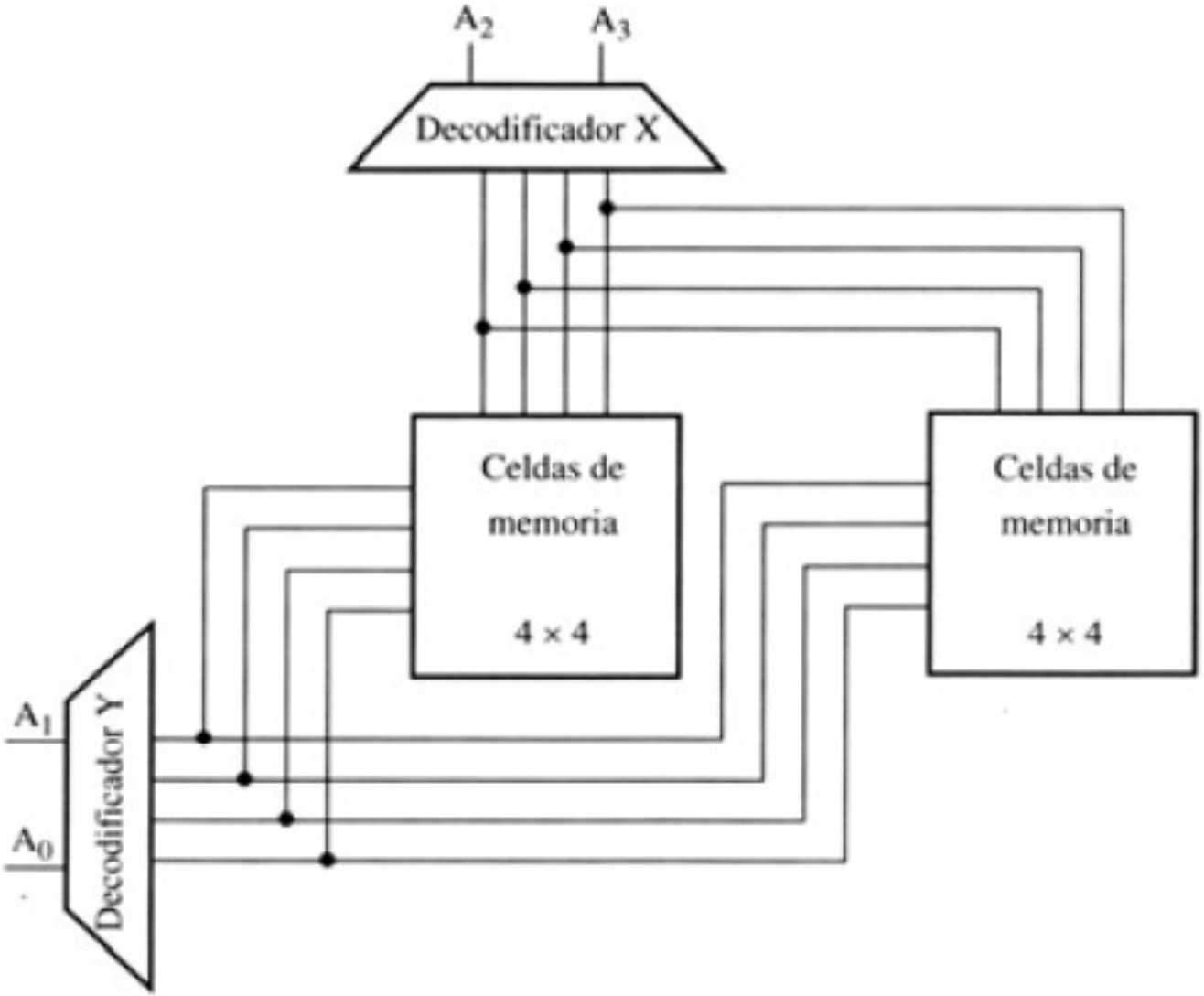
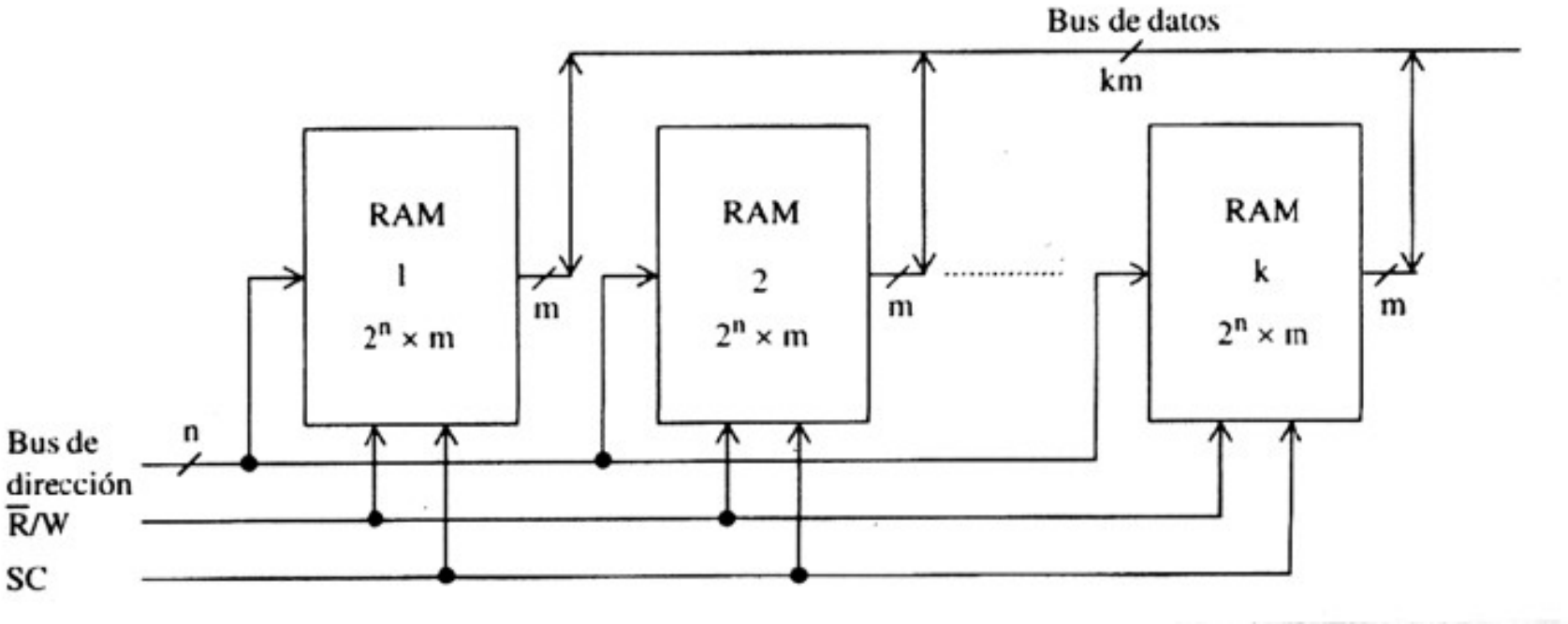
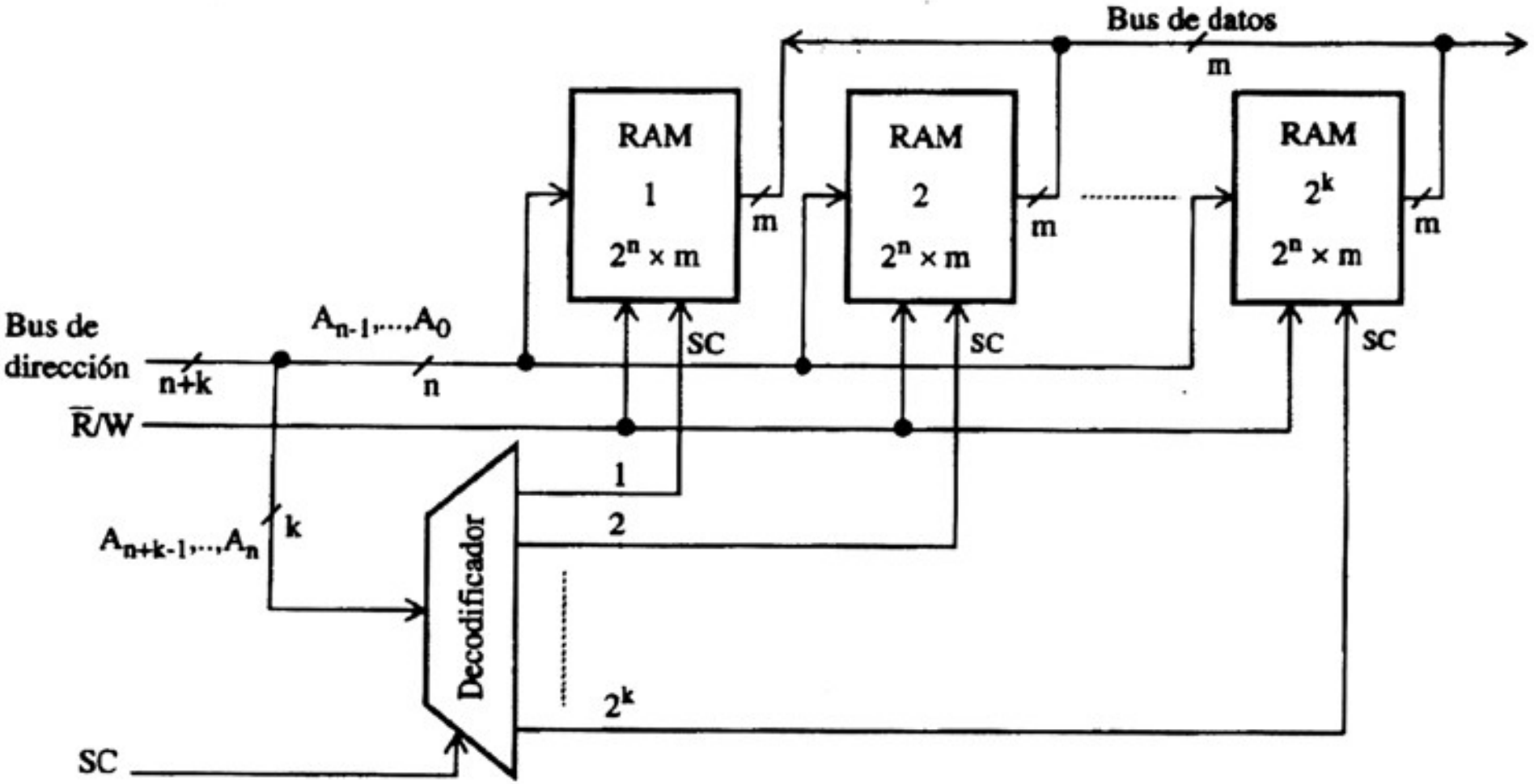
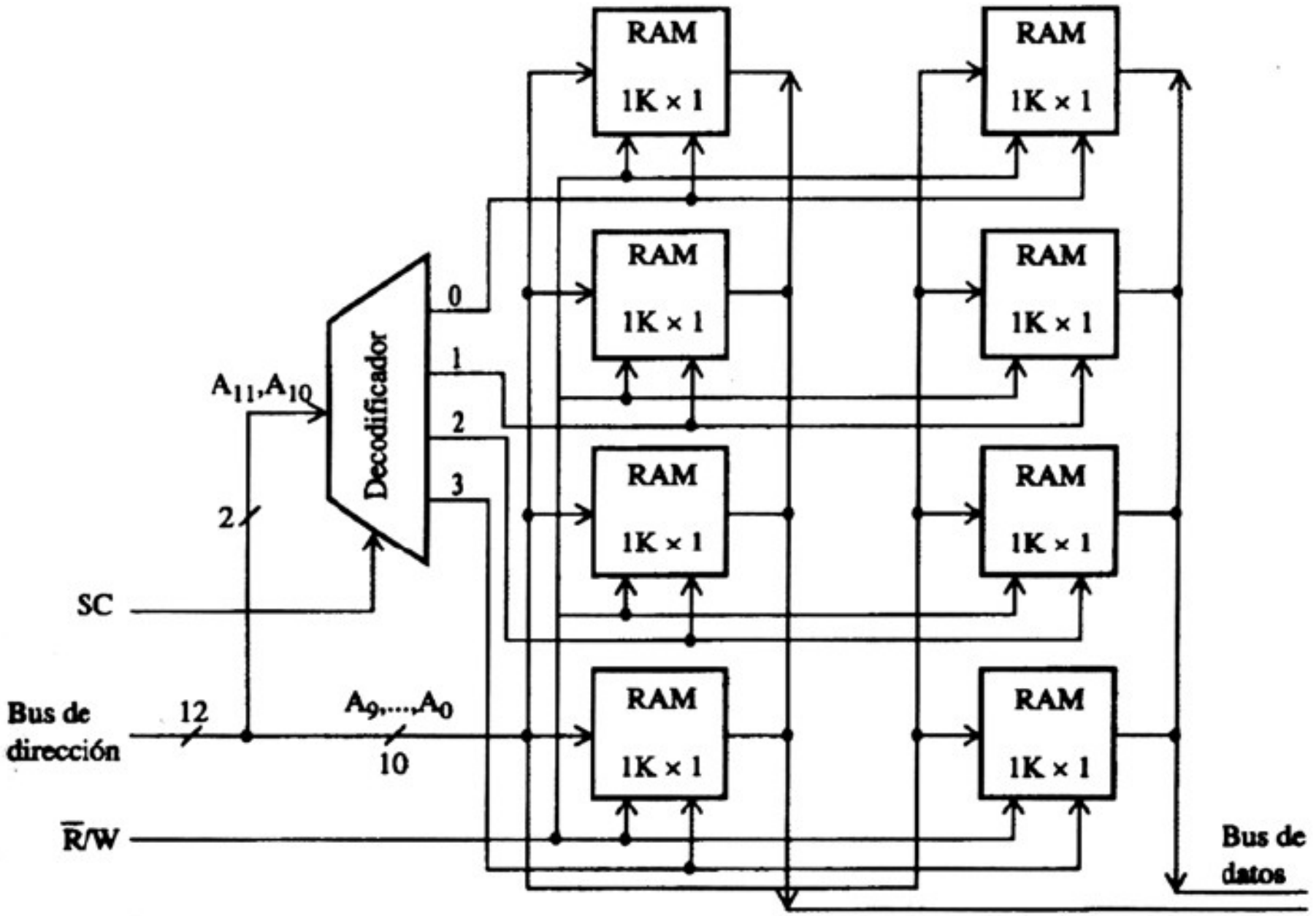


Figura 2.19: Memoria RAM de 16 × 2 con selección por coincidencia

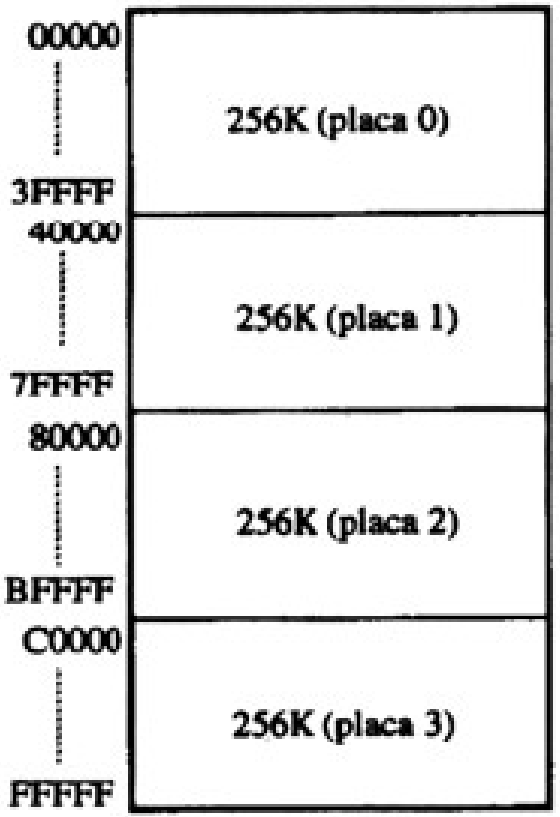
Diseños de bloques de memoria




Bus de datos
m
RAM
1
$2^n \times m$
RAM
2
$2^n \times m$
RAM
2^k
$2^n \times m$
m
m
m
Bus de
dirección
$A_{n-1},\ldots,A_0$
n+k
n
SC
SC
SC
$\overline{R}/W$
1
2
2^k
$A_{n+k-1},\ldots,A_n$
k
Decodificador
SC


RAM
1K × 1
RAM
1K × 1
RAM
1K × 1
RAM
1K × 1
RAM
1K × 1
RAM
1K × 1
RAM
1K × 1
RAM
1K × 1
Decodificador
0
1
2
3
A11,A10
2
SC
Bus de
dirección
12
A9,....,A0
10
R/W
Bus de
datos

Mapa de memoria



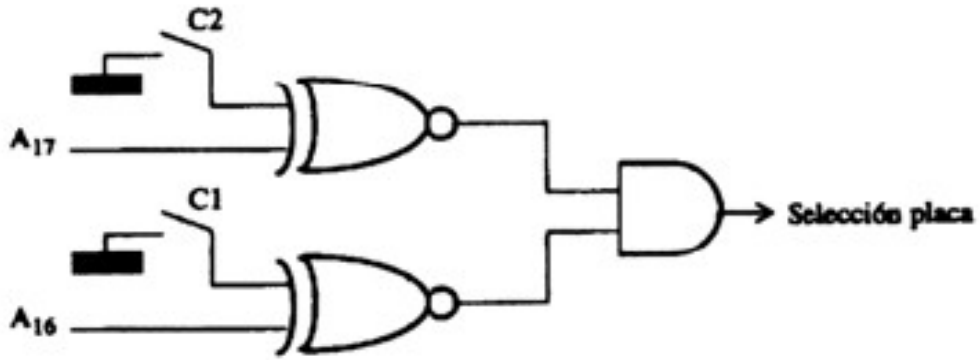
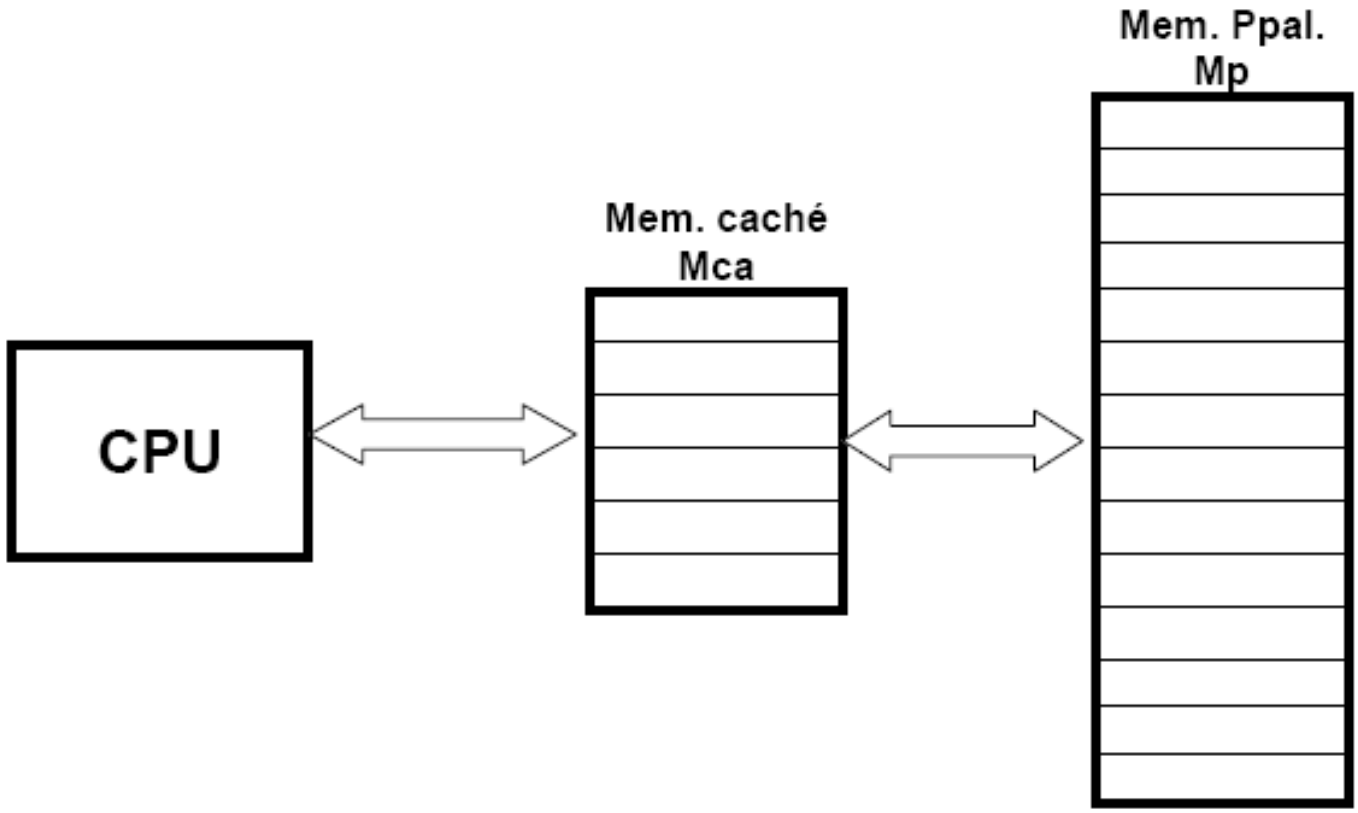
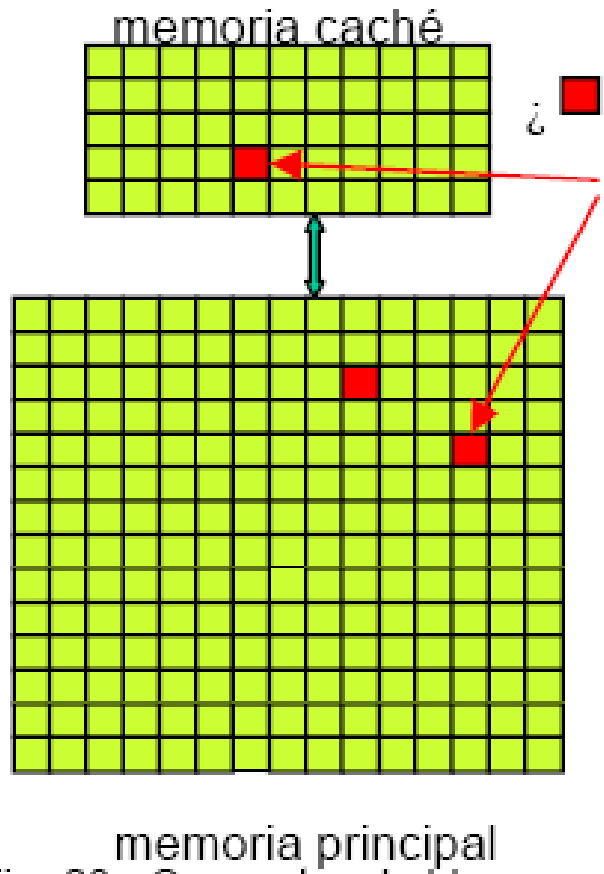


Figura 2.29: Mapa de memoria

2.4 Memorias caché

- Para solucionar el compromiso entre velocidad, coste y capacidad se coloca una memoria pequeña y rápida entre la UCP y la memoria principal, la memoria caché.
- La memoria caché almacena una copia de ciertas partes de la memoria principal.
- Cuando la UCP intenta leer una palabra en primer lugar comprueba si está en la caché.
  - Si está, se lee,
  - Si no está, se transfiere a la memoria caché un bloque de la memoria principal, con un determinado número de palabras.
- **Acierto**
  - Cuando el dato solicitado por la CPU está en la Mca
- **Fallo**
  - Cuando el dato solicitado por la CPU no está en la Mca


Mem. Ppal.
Mp
Mem. caché
Mca
CPU



Bloque:
Cantidad mínima de información que puede estar presente o no en M_p y M_{ca}

Acierto:
Cuando el dato solicitado está en la M_{ca}

Fallo:
Cuando el dato solicitado no está en la M_{ca}

Fig. 03. Conceptos de bloque, acierto y fallo de una memoria caché.

Organización

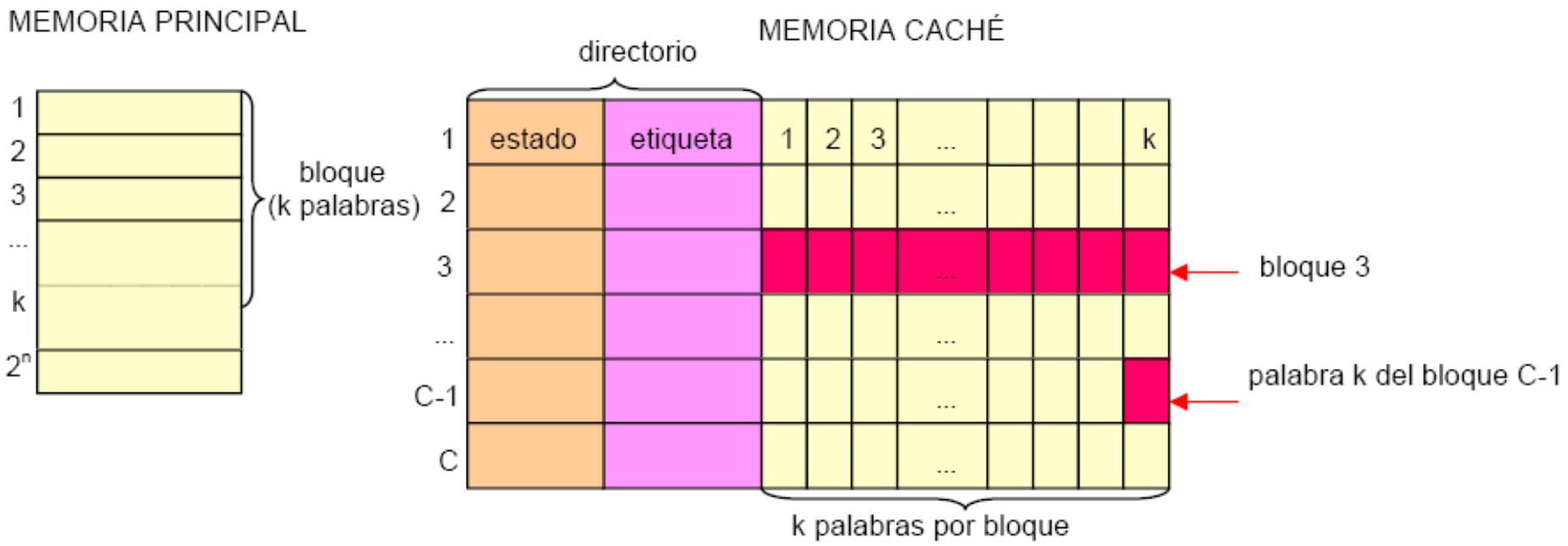
- La memoria principal consta de 2^n palabras y cada palabra se puede referenciar mediante una dirección única de n bits
- Para realizar la transformación entre la memoria principal y la caché, se considera que la memoria principal está constituida por una serie de bloques de longitud fija de k palabras/bloque, es decir, hay M = 2^n/ k bloques.
- La memoria caché contiene C particiones de k palabras (C << M).

Mp $\Rightarrow$ n bits en el bus de dir $\Rightarrow$ 2^n palabras

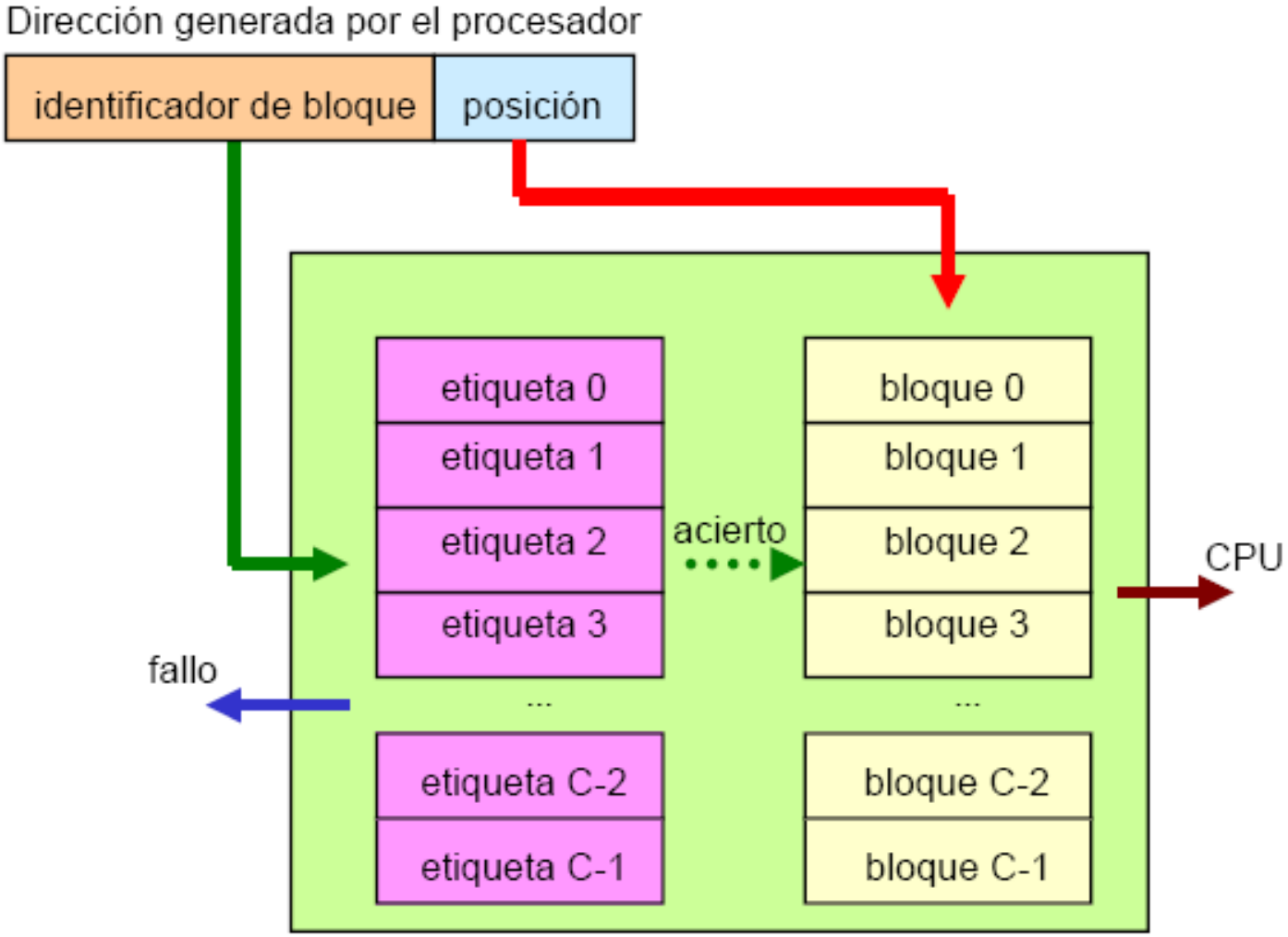
M bloques de k palabras por bloque $\Rightarrow$ $M = \frac{2^n}{k}$

Mca $\Rightarrow$ c bloques de k palabras cada uno y en cada dirección una etiqueta indicativa de la dirección

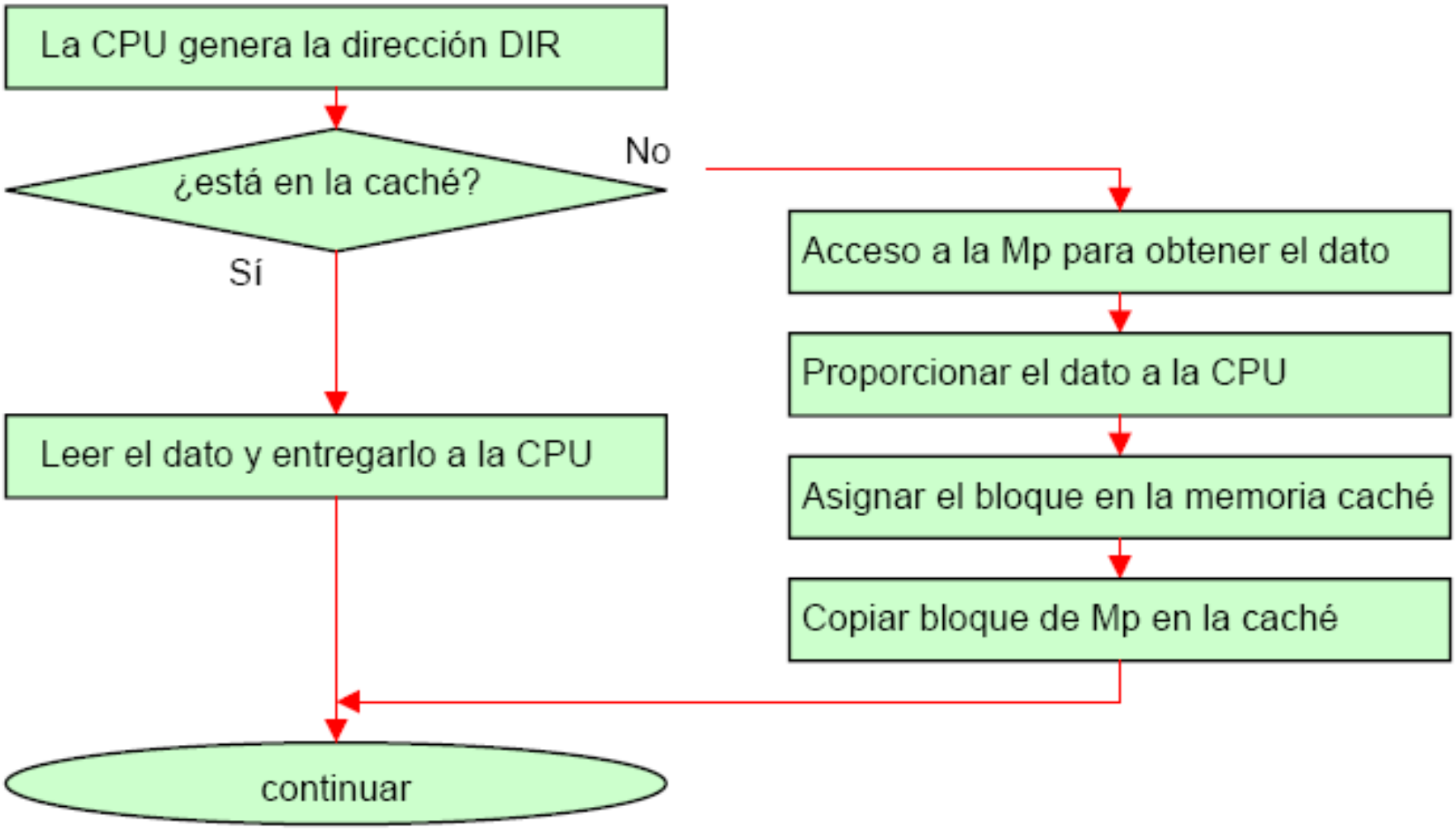
M >> c


MEMORIA PRINCIPAL
1
2
3
...
k
2^n
bloque
(k palabras)
MEMORIA CACHÉ
directorio
estado
etiqueta
1
2
3
...
k
1
2
3
...
C-1
C
bloque 3
palabra k del bloque C-1
k palabras por bloque

Acceso a la memoria caché en operación de lectura



Operación de lectura en Memoria Cache



Criterios de diseño	- Capacidad	1K, 4K, 16K, 32K
	- Organización	Directa Totalmente asociativa Asociativa por conjuntos
	- Mecanismo de búsqueda	Por demanda Con anticipación Selectiva
	- Algoritmo reemplazamiento	Utilizada menos frecuentemente (LRU) Más antigua (FIFO) Utilizada menos frecuentemente (LFU) Aleatorio
	- Estrategia escritura	Escritura inmediata Post-escritura Escritura única
	- Tamaño de bloque	4, 8, 16, 32.. palabras
	- Número de cachés	Número de niveles

Rendimiento de la memoria caché

- El rendimiento de una memoria caché se mide en la tasa de acierto.
  - Tasa de acierto = Acierto/(Acierto + Fallo)

Tasa de acierto $h = \frac{aciertos}{aciertos + fallos} = \frac{aciertos}{accesos} \Rightarrow h \geq 0{,}9 \Rightarrow$ ***principio de localidad***

Tasa de fallos = 1-h

Tiempo acceso medio $ta = h \times tca + (1 - h) \times tp$

tca= t. Acces. Medio a Mca
tp = t. Acces. Medio a Mp

(tau) $\tau = \frac{tca}{tp} \Rightarrow 0{,}1 \leq \tau \leq 0{,}5$

lambda = índice de mejora = $\lambda = \frac{tp}{ta} = \frac{1}{1 - h \times (1 - \tau)}$

Índice de mejora λ de la Mca en función de la tasa de acierto h

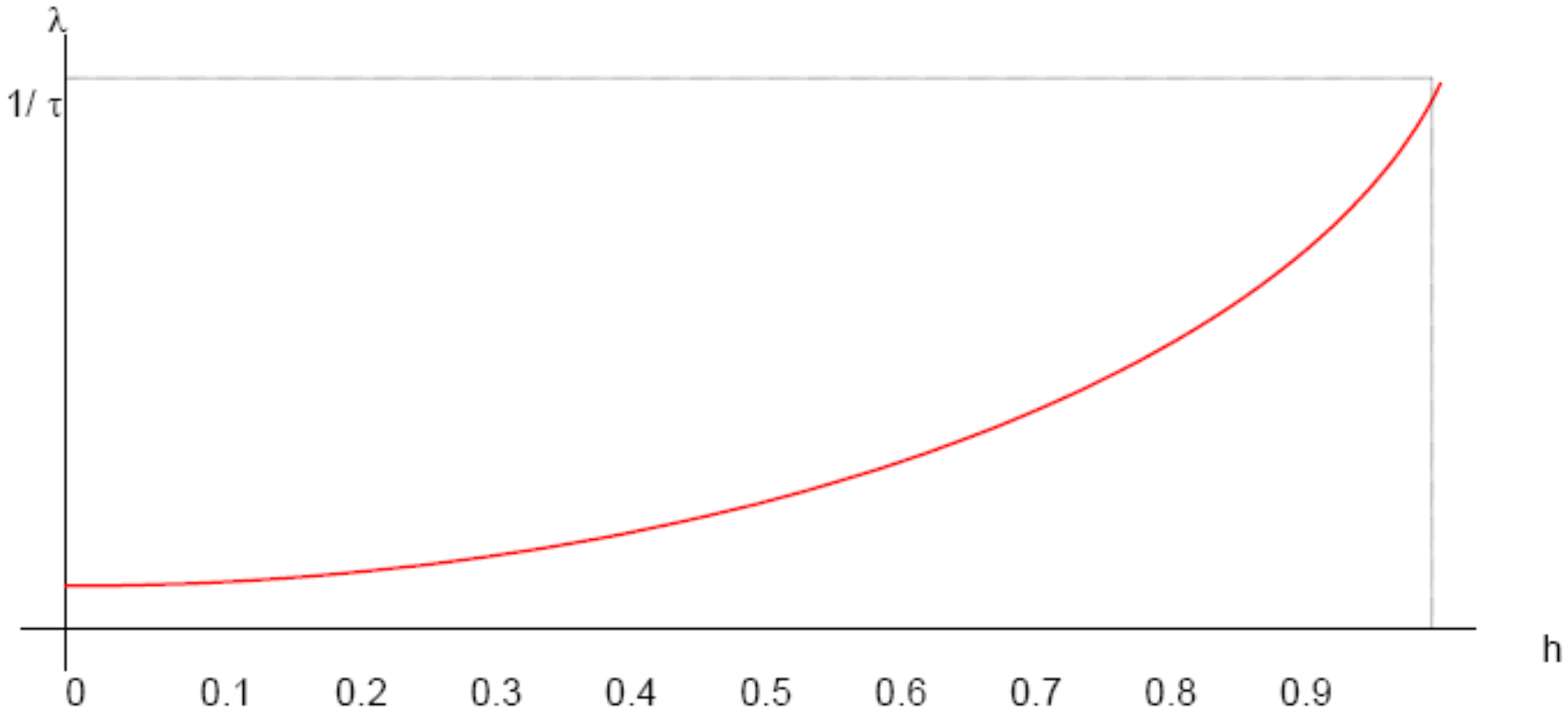


Fig. 10. Índice de mejora λ de la M_{ca} en función de la tasa de acierto h.

Capacidad de la caché

- El tamaño de la caché plantea el compromiso de ser pequeña para disminuir el coste medio por bit almacenado en la memoria interna (caché + principal), por otro lado debe ser grande como para que tenga una tasa de acierto elevada.

Organización de la memoria caché

- Establecer la función de correspondencia que asigna a los bloques de la memoria principal en las posiciones definidas en la memoria caché
- Técnicas
  - Directa
  - Totalmente asociativa
  - Asociativa por conjuntos

Parámetros del ejemplo a utilizar en las descripciones

- a) Ancho de palabra de datos; p.ej. 16 bits
- b) Tamaño de la Memoria caché; p.ej.512 B = 2^9 Bytes
- c) Tamaño de bloque; k = 8
- d) Tamaño de la Memoria principal = 32 KB
- Ejemplos:
  - 32 KB = 2^{15} ⇒ Bus de direcciones = 15 bits ⇒ A0 a A14
  - 512B = 2^9 ⇒ Bus direcciones de la caché ⇒ 9 bits
  - 512B y k = 8 -> Nº bloques = 512/8 = 64 bloques
  - k = 8 ->2^3 → 3 bits
  - 64 bloques = 2^6 → 6 bits

Correspondencia directa

- Cada bloque de memoria principal se transforma en una única partición de la memoria caché.

Principios:

- Palabras por bloque 8 ⇒ 3 bits (b0 a b2)
- Nº bloques 64 ⇒ 6 bits
- Etiqueta = (bus direcc) – (bits de bloques) -- (bits palab/bloque) ⇒ 15 – 6 – 3 = 6
- Ancho de la mem. caché = Ancho palabra + ancho etiqueta ⇒ 16 + 6 = 22 bits

Formato de instrucción

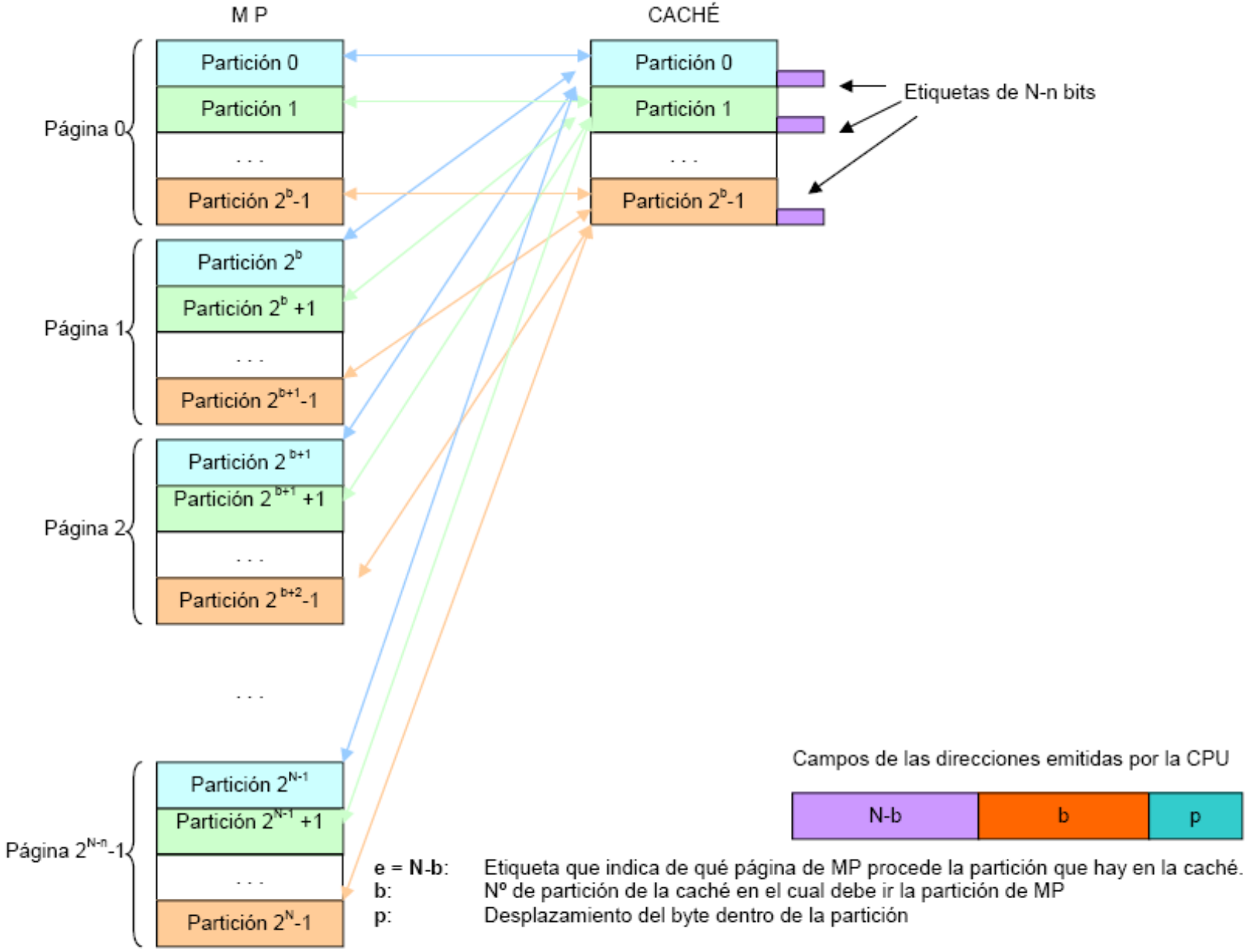


Fig. 12. Asignación de bloques de la M_P en la M_{ca} con correspondencia directa.

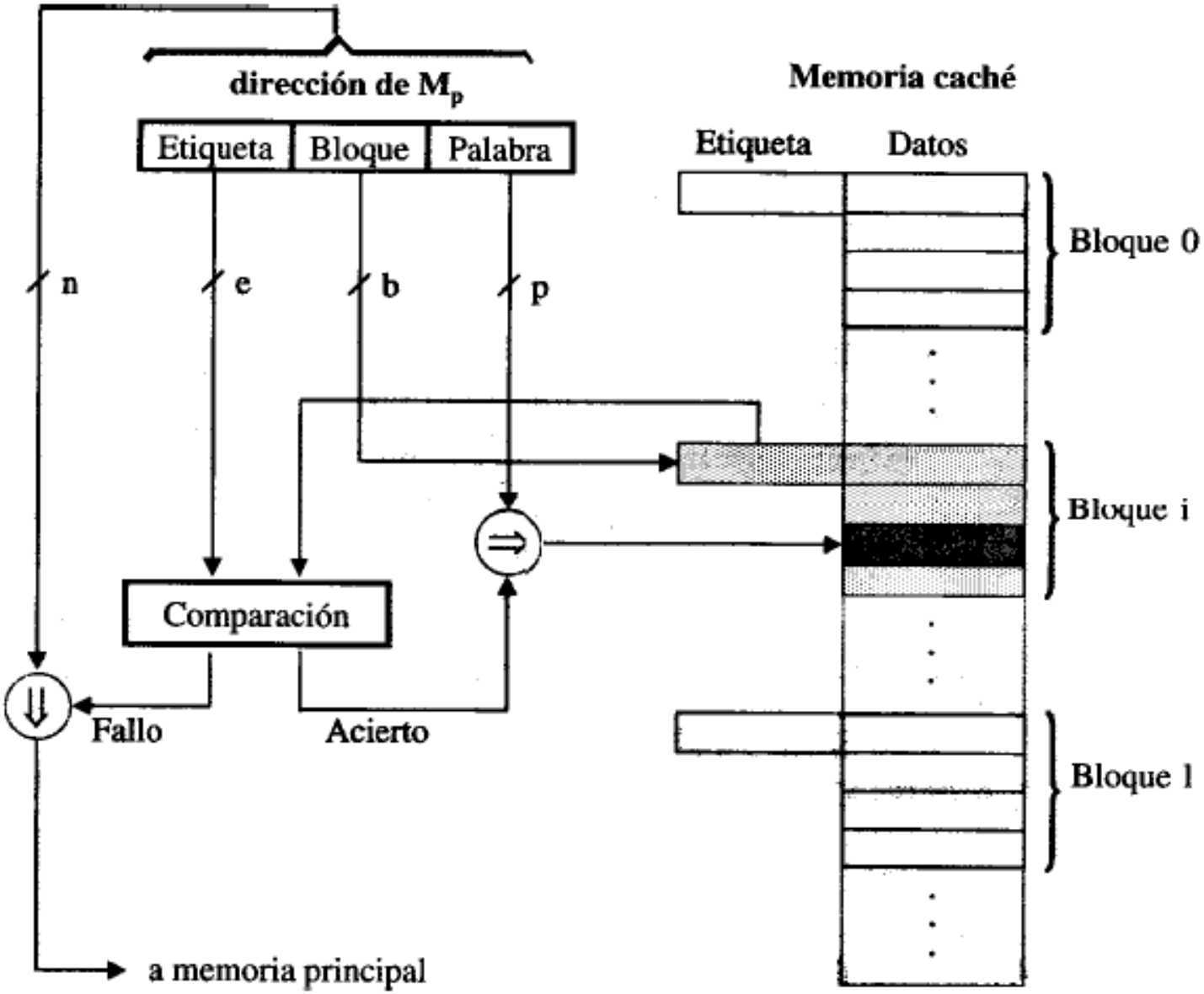


Fig. 13. Organización de una memoria caché con correspondencia directa.

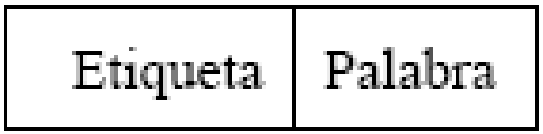
Correspondencia totalmente asociativa

- Permite que se cargue un bloque de memoria principal en cualquier partición de la memoria caché
- Este sistema permite decidir cual es el bloque que será sustituido en la caché por el nuevo bloque leído en memoria principal.
- Su principal desventaja es la necesidad de una circuitería compleja para examinar en paralelo los campos de etiqueta de todas las particiones de la memoria caché. Formato de instrucción

Principios:

- Palabras por bloque 8 $\Rightarrow$ 3 bits (b0 a b2)
- Nº bloques 64
- Etiqueta = (bus direcc) – (bits palab/bloque) $\Rightarrow$ 15 –3 = 12
- Ancho de la mem. caché = Ancho palabra + ancho etiqueta $\Rightarrow$ 16 + 12 = 28 bits

Formato de instrucción



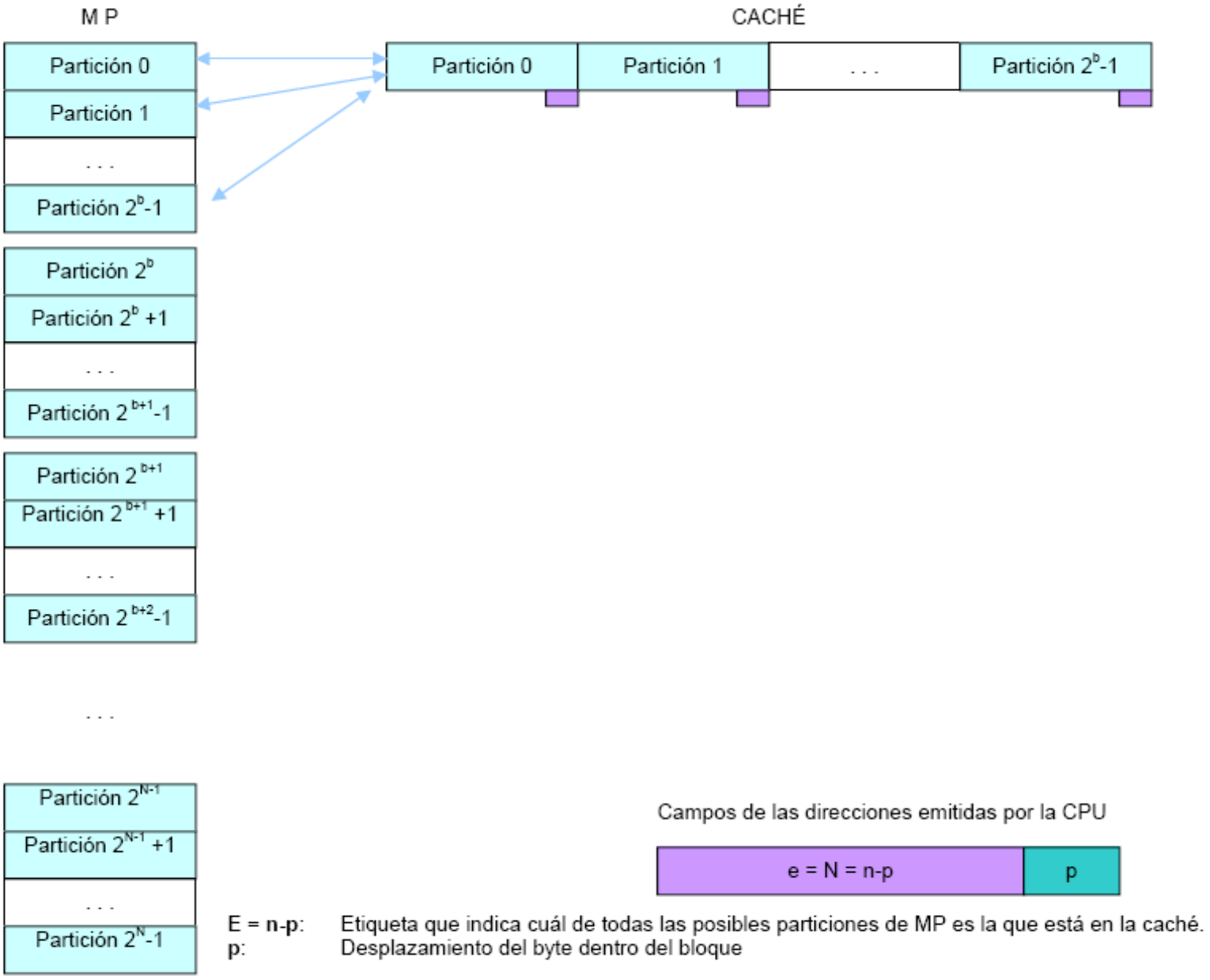


Fig. 15. Asignación de bloques de la M_P en la M_{ca} con correspondencia totalmente asociativa.

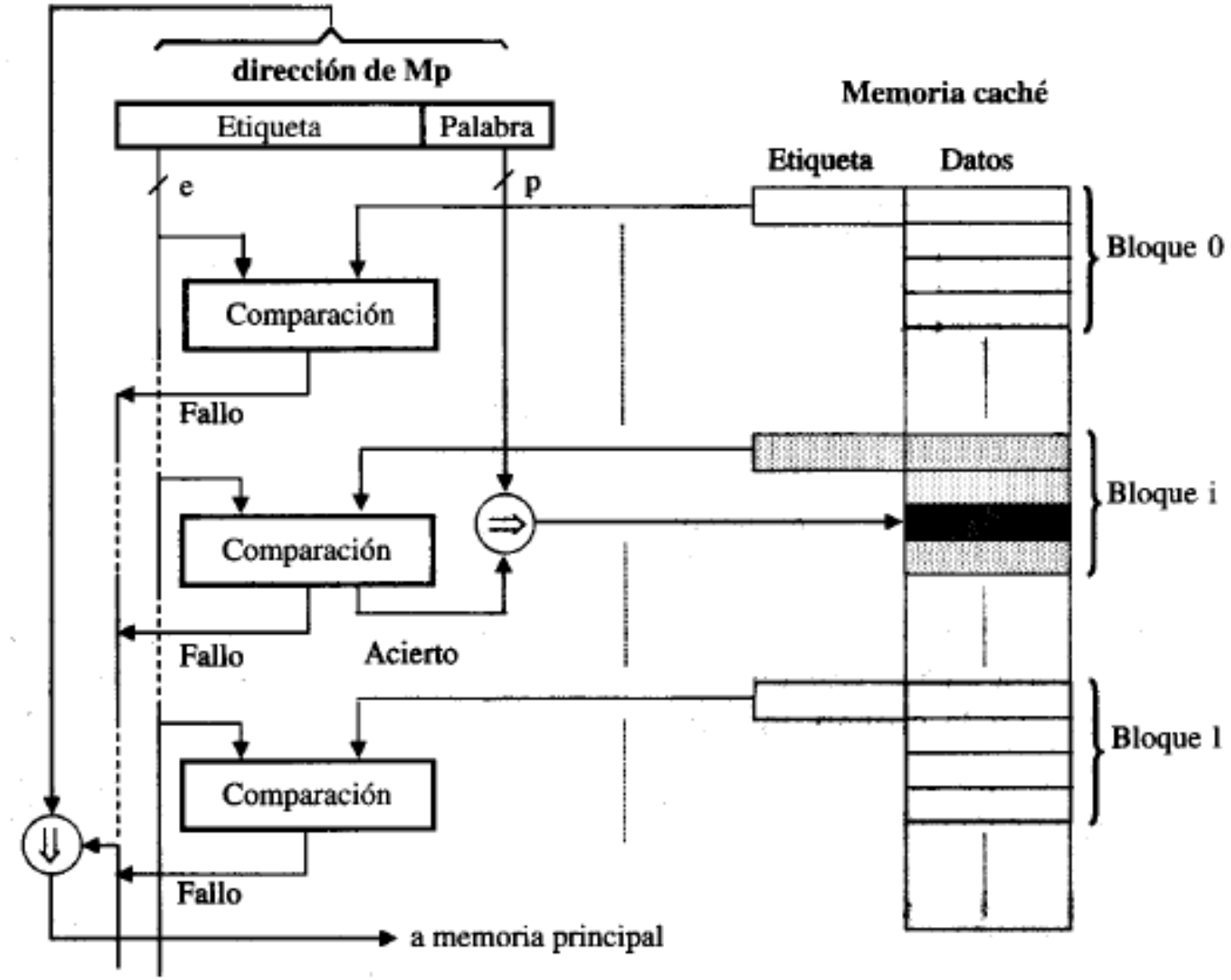


Fig. 16. Organización de una memoria caché con correspondencia totalmente asociativa.

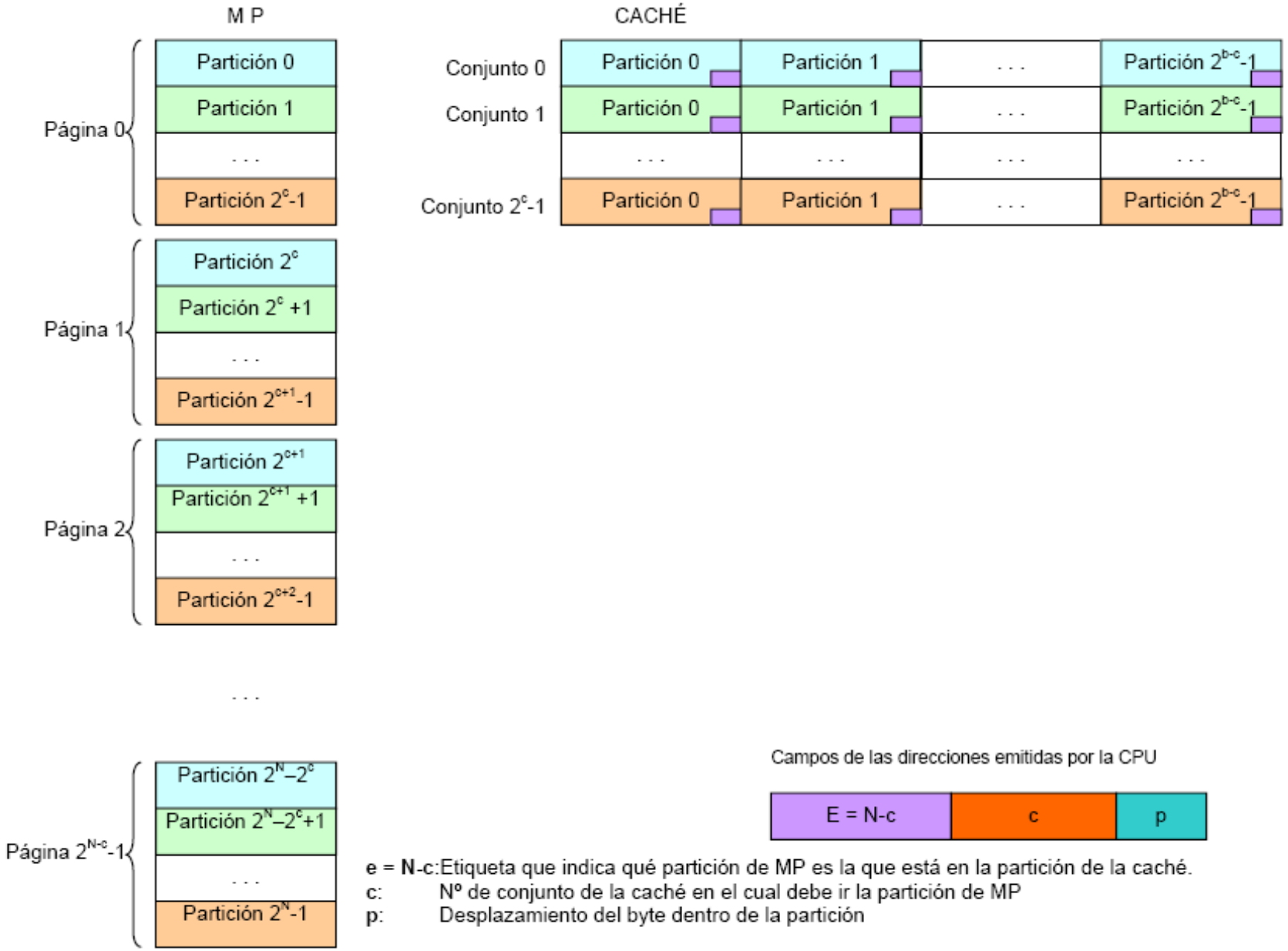
Asociativa por conjunto

- Auna las ventajas de los dos métodos anteriores.
- Está compuesta por “r” bloques y “q” conjuntos de modo que C = q×r , siendo C el nº de bloques de la mem. caché.
- El funcionamiento consiste en que cada bloque de la mem. ppal. tiene asignado un conjunto de la caché, pero se puede ubicar en cualquiera de los bloques que pertenecen a dicho conjunto. Ello permite mayor flexibilidad que la correspondencia directa y menor cantidad de comparaciones que la totalmente asociativa.

Principios:

- Palabras por bloque 8 $\Rightarrow$ 3 bits (b0 a b2)
- Nº bloques/conjunto =2 $\Rightarrow$ Nº conjuntos 64 /2 = 32 $\Rightarrow$ 2^5 $\Rightarrow$ 5 bits
- Etiqueta = (bus direcc) – (bits palab/bloque) -- (bits conjuntos) $\Rightarrow$ 15 –3 – 5 = 7
- Ancho de la mem. caché = Ancho palabra + ancho etiqueta $\Rightarrow$ 16 + 7 = 23 bits

Formato de instrucción



e = N-c: Etiqueta que indica qué partición de MP es la que está en la partición de la caché.
c: Nº de conjunto de la caché en el cual debe ir la partición de MP
p: Desplazamiento del byte dentro de la partición

Fig. 18. Asignación de bloques de la M_P en la Mca con correspondencia asociativa por conjuntos.

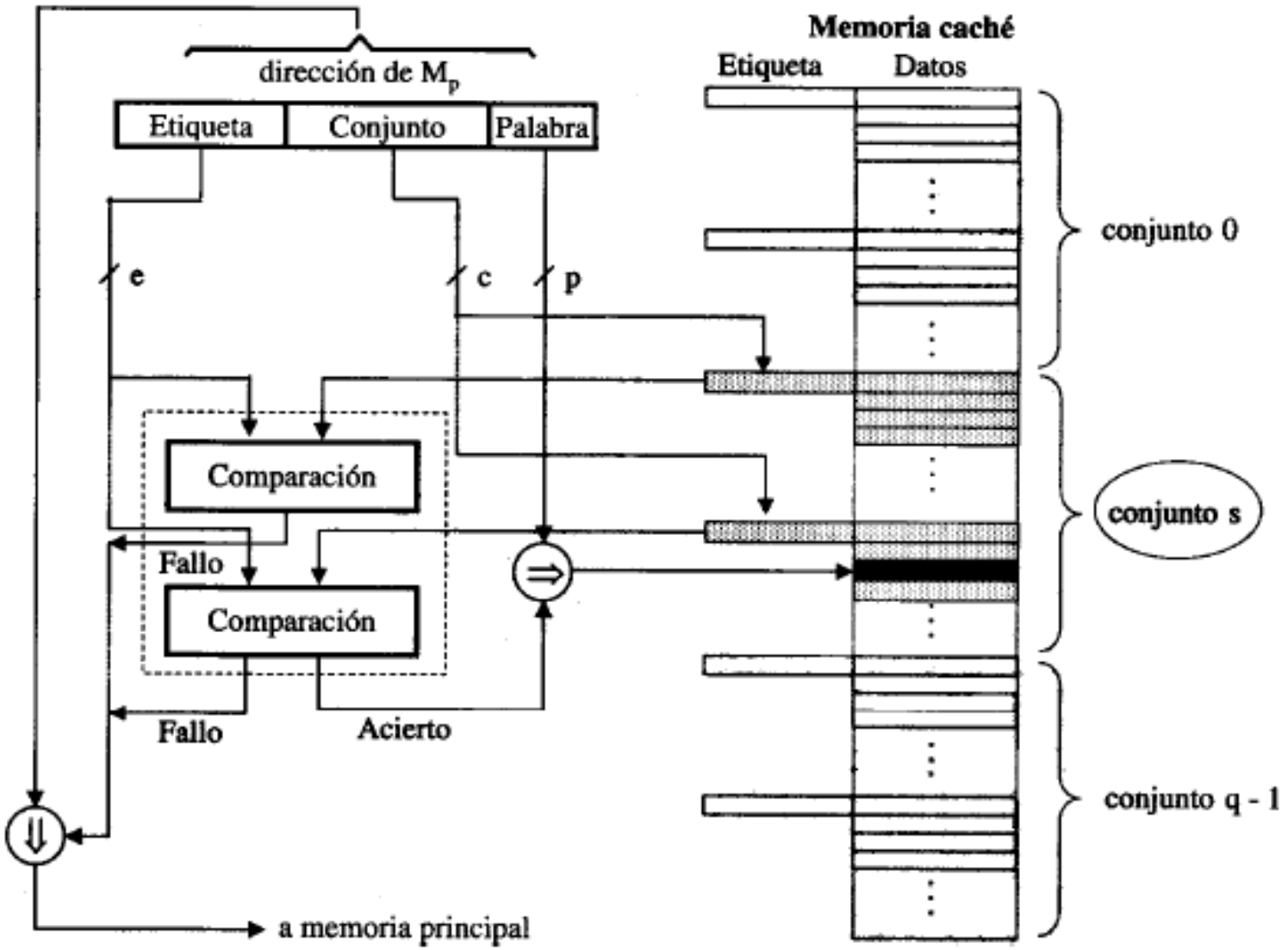


Fig. 19. Organización de una memoria caché con correspondencia asociativa por conjuntos.

Algoritmos de reemplazamiento

- Necesarios en asociativo y conjunto
- Partición más antigua en la memoria caché sin ser referenciada (LRU).
- Partición más antigua en la memoria caché (FIFO).
- Partición utilizada menos frecuentemente. (LFU).
- Partición elegida de forma aleatoria.

Estrategia de escritura

- Se plantean dos problemas:
  - Más de un dispositivo tiene acceso a la memoria principal.
    - Si la palabra ha sido modificada en la memoria caché el dato de la memoria principal no es válido, si por el contrario ha sido modificado en memoria principal el dato de la caché no es válido.
  - Se conectan varias UPC's a un bus y cada una tiene su propia memoria caché local.

Escritura Directa

- Efectuar todas las operaciones de escritura tanto en memoria principal como en memoria caché
- Su desventaja
  - Que genera un tráfico elevado con la memoria y puede crear un cuello de botella

Postescritura

- Cuando se produce una modificación se pone a “1” un bit de actualización asociado con cada partición de la memoria caché.
- Si se reemplaza un bloque se reescribe la memoria principal si y sólo si el bit de actualización está a “1”.
- El problema es saber que partes de la memoria principal ya no son válidas, para evitarlo los accesos de los módulos de E/S se permiten únicamente a través de la caché.

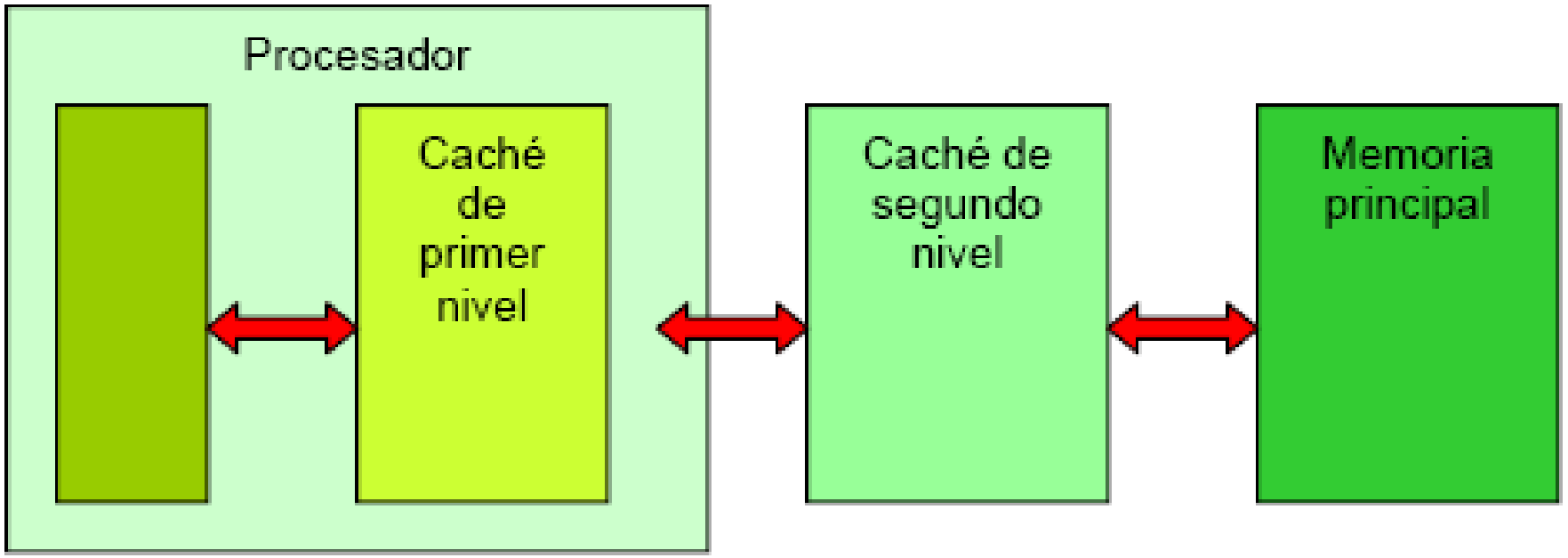


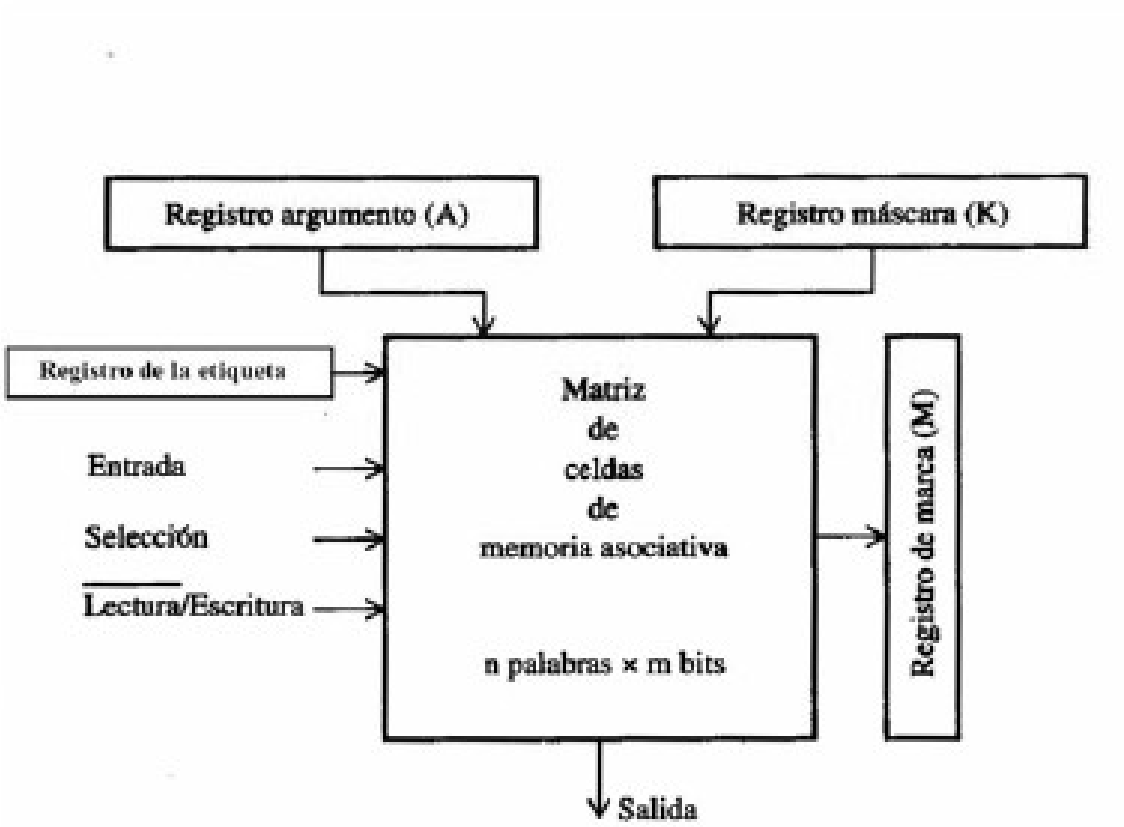
Fig. 29. Memoria caché de dos niveles.

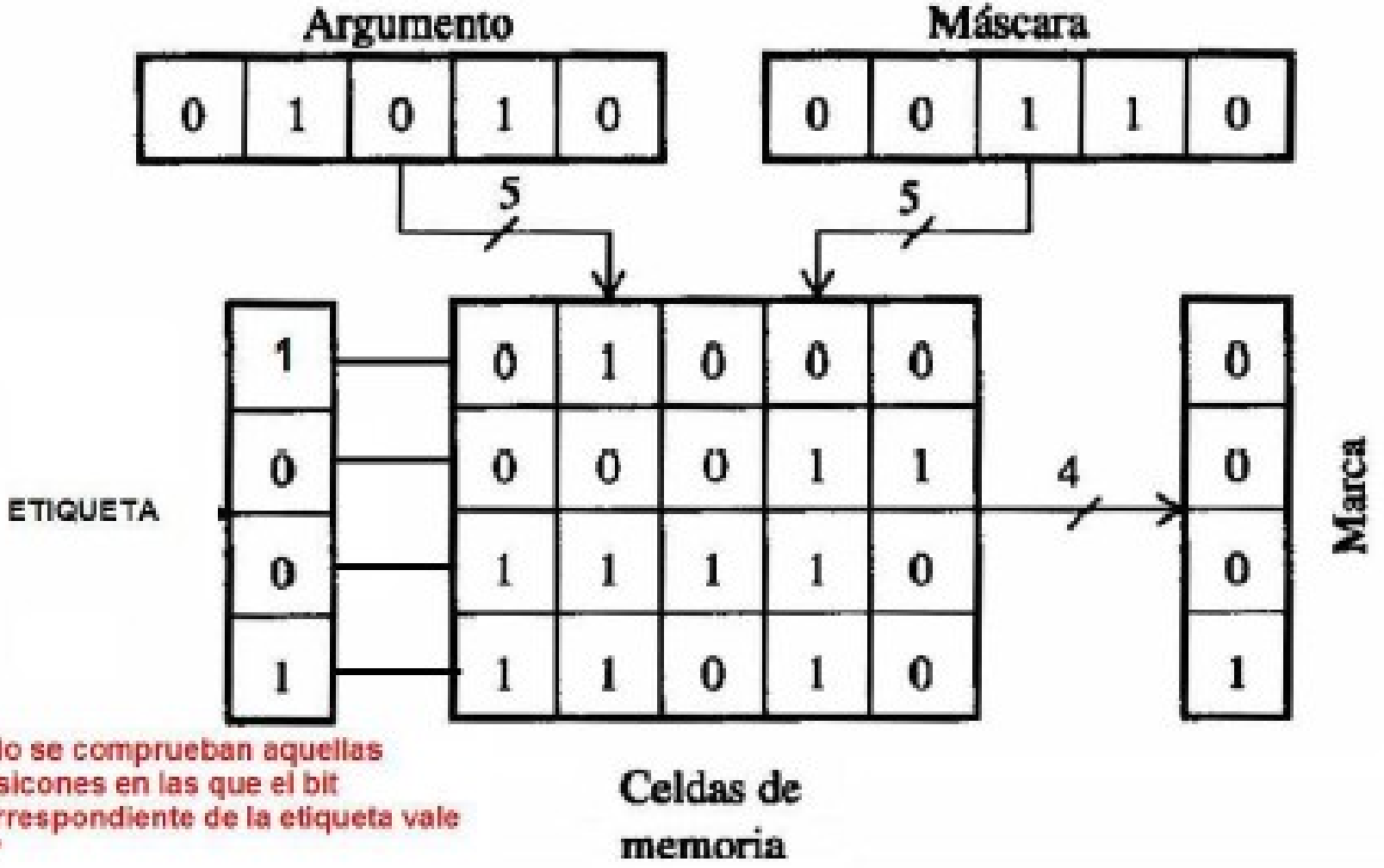
2.5 Memorias asociativas

- Búsqueda por contenido
- Consiste en una matriz de celdas de memoria, con su lógica asociada, organizada en **n** palabras con **m** bits/palabra.
- El registro argumento **A** y el registro de máscara **k** tienen **m** bits cada uno y el registro de marca **M** consta de **n** bits.
- Cada palabra de la memoria se compara en paralelo con el contenido del registro argumento, y se pone a “1” el bit del registro de marca asociado a aquellas palabras cuyo contenido coincide con el del registro argumento.
- El proceso de lectura se realiza mediante un acceso secuencial a las palabras de memoria cuyos bits en el registro de marca están a “1”
- El registro de máscara proporciona la clave para seleccionar la parte que se desee en la palabra argumento.

Estructura

- Matriz de celdas de memoria
- Registro argumento
- Registro máscara
- Registro de marca
- Registro de la etiqueta





Solo se comprueban aquellas posicones en las que el bit correspondiente de la etiqueta vale "1"

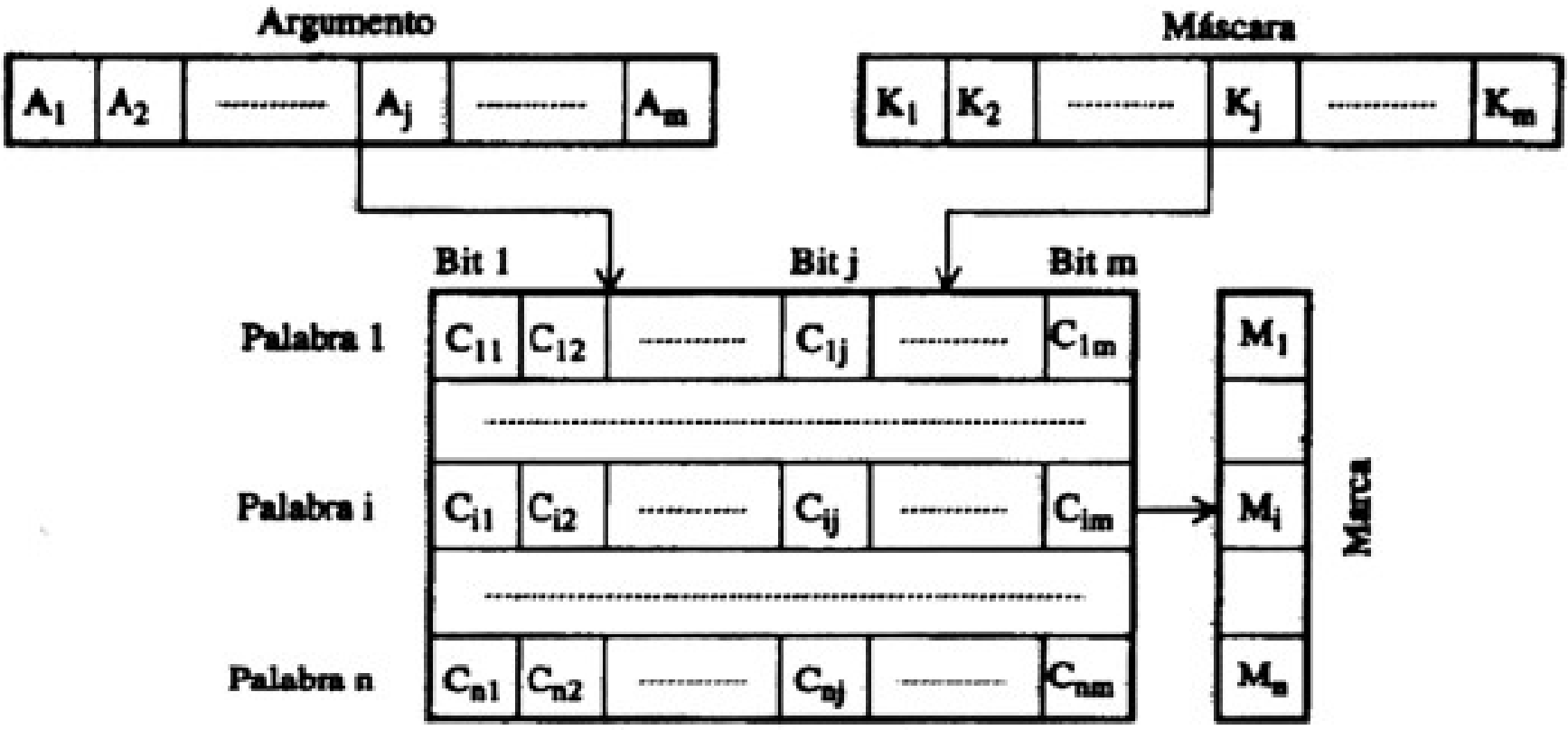
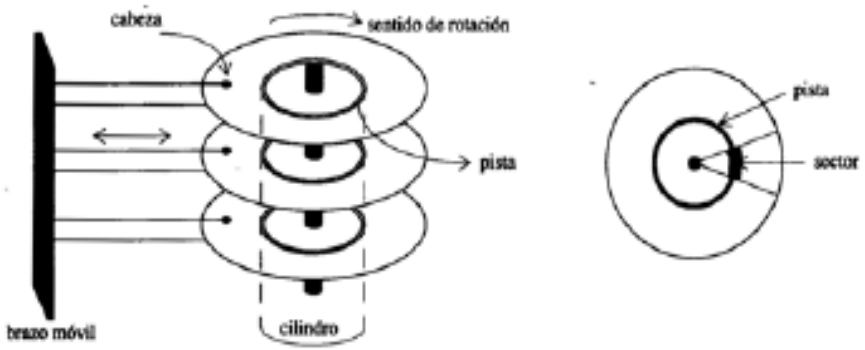


Figura 2.51: Memoria asociativa de n palabras × m bits

2.8 Discos magnéticos

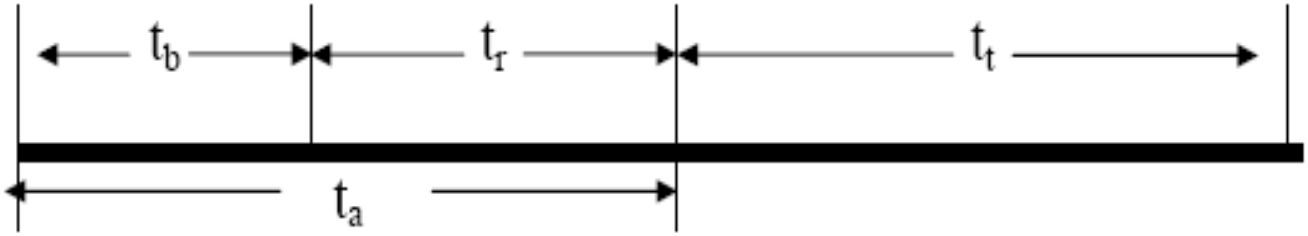
Estructura Física ⇒ Película de óxido magnético sobre soporte inerte (aluminio o plástico)

Estructura
- Cabezas de lectura/escritura ⇒ Una por cara
- Cada disco dos superficies
- Pistas concéntricas
- Cilindro ⇒ conjunto de pistas paralelas de todas las superficies
- Sector ⇒ porción continuada de visten que se divide cada pista

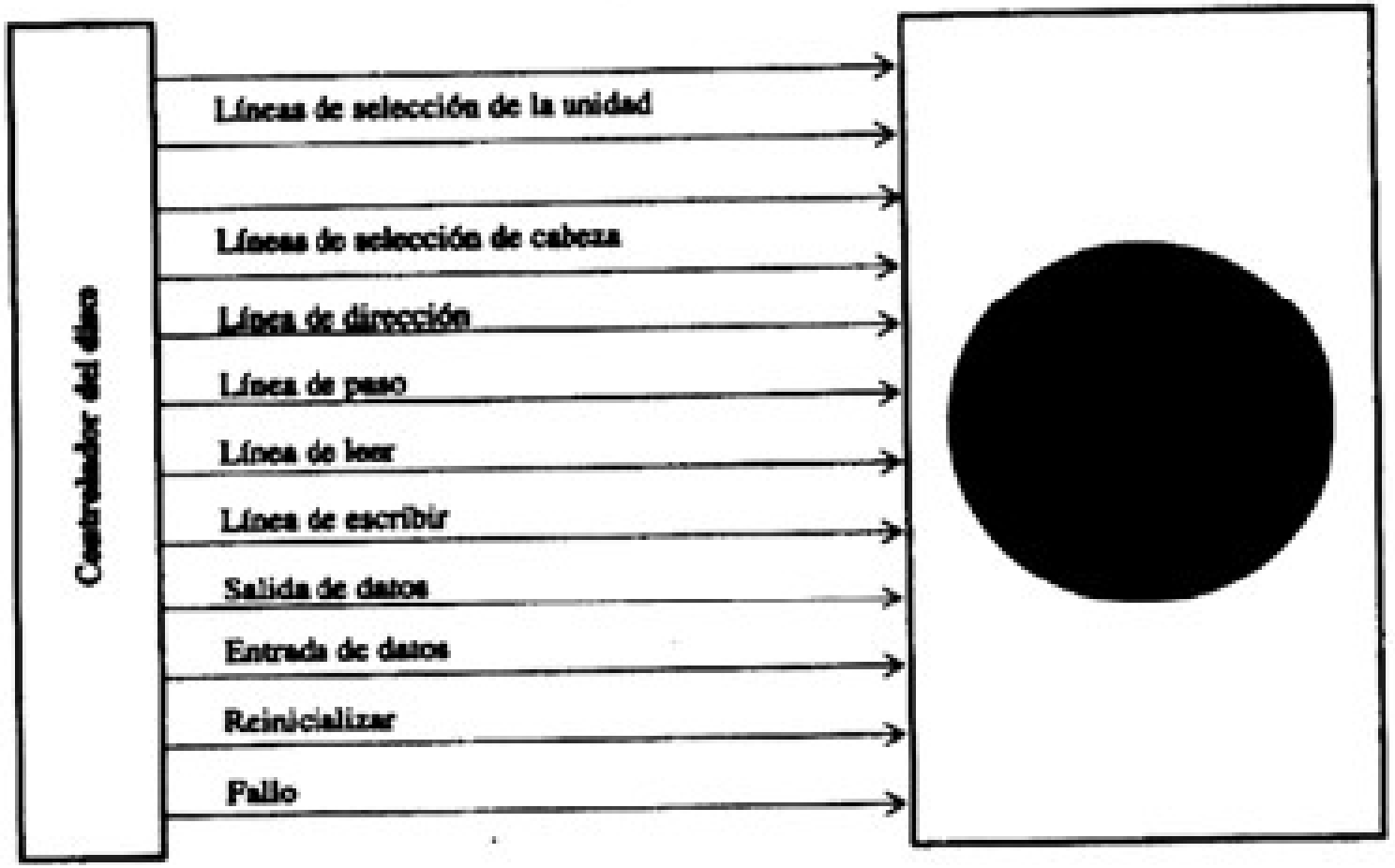


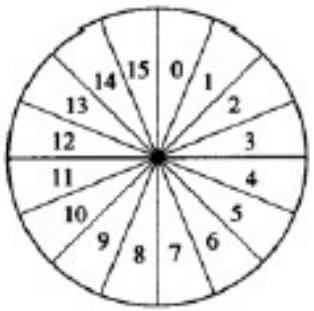
PRINCIPIOS DE CONSTITUCIÓN Y FUNCIONAMIENTO		
Velocidad de giro constante	⇒	Densidad de grabación diferente en las diferentes pistas
Nº sectores/pista constante en todas las pistas		
Nº Bytes/sector constante		
Pistas tienen diferente radio		

TIEMPOS			
Tiempo de búsqueda (t_b)	Posicionamiento de la cabeza en el cilindro	$t_b = m \times n + t_i$	*m= cte del disco* *n=n° cil. Desplaz.*
Tiempo de latencia rotacional (t_r)	Girar disco y posicionar la cabeza en el sector	$t_r = \frac{1}{2 \times f}$	*f = veloc. Rotac.*
Tiempo de acceso	$ta = t_b + t_r$		
Tiempo de transferencia (t_t)	Transferencia de datos una vez posicionada la cabeza	$t_t = \frac{b}{P \times f}$	*b =n° byte a transfe.* *P = n° bytes/pista*

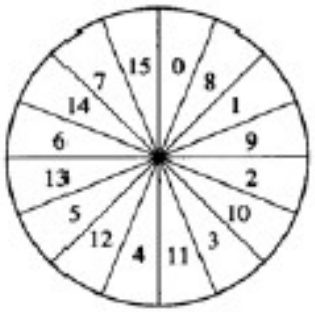


Controlador de disco

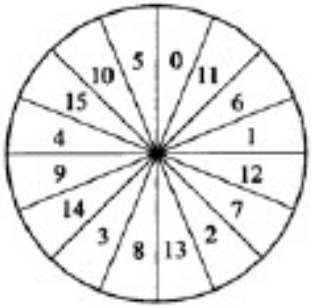




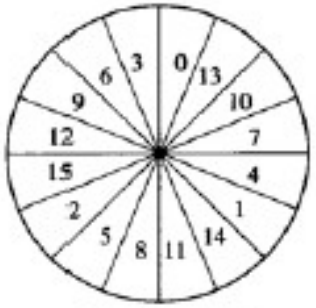
sin entrelazado



entrelazado simple



entrelazado doble



entrelazado cuadruple

ENTRELAZADO: Distribución no consecutiva de sectores que permite tratar los errores después de la lectura de cada sector dando tiempo a leer el siguiente sector lógico sin tener que dar una vuelta completa el disco.
Depende de la velocidad de giro del disco y del controlador.

PLANIFICACIÓN DEL DISCO: Forma de recorrer los sectores de un disco cuando se dispone de una lista de sectores a los cuales acceder.

Planificaciones de acceso a los sectores

- **FCFS** — First Come First Secued ⇒ FIFO ⇒ 1º entra, 1º sale
- **SSTF** — Shortest Service Time First ⇒ 1º el más cercano
- **SCAN** — Rastreo ⇒ todas pistas en una dirección u otra.
- **C-SCAN** — Una única dirección.
- **LOOK/C-LOOK** — Igual a SCAN pero sin llegar al fin.

Orden de peticiones: 22, 124, 105, 181, 142, 36, 5, 59, 115.
Posición inicial: 95

Próxima pista a la que se accede	22	124	105	181	142	36	5	59	115	
Número de pistas que se atraviesan	73	102	19	76	39	106	31	54	56	LMB = 61,8

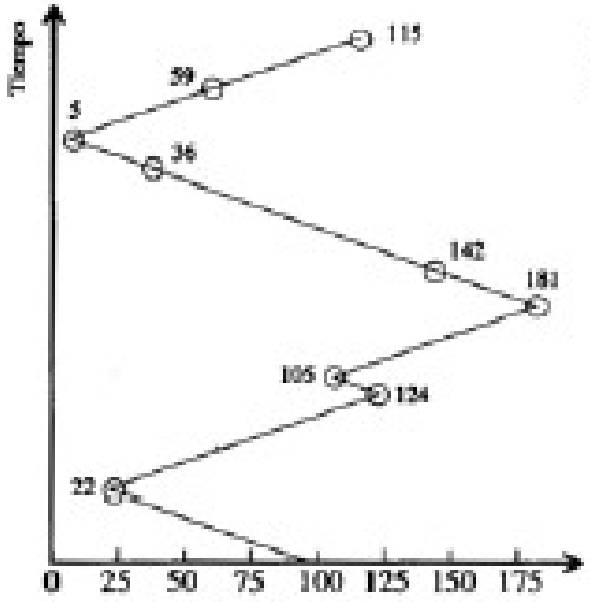
Tabla 2.18: Algoritmo FCFS de planificación del disco

Próxima pista a la que se accede	105	115	124	142	181	59	36	22	5	
Número de pistas que se atraviesan	10	10	9	18	39	122	23	14	17	LMB = 29,1

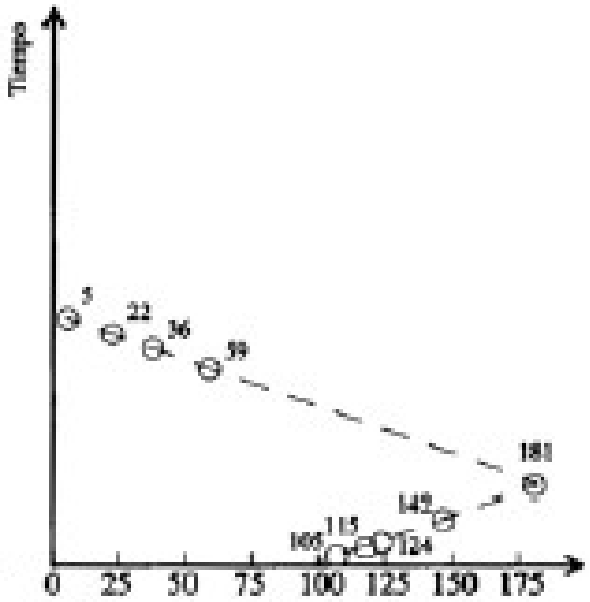
Tabla 2.19: Algoritmo SSTF de planificación del disco

Próxima pista a la que se accede	59	36	22	5	105	115	124	142	181	
Número de pistas que se atraviesan	36	23	14	17	100	10	9	18	39	LMB = 29,5

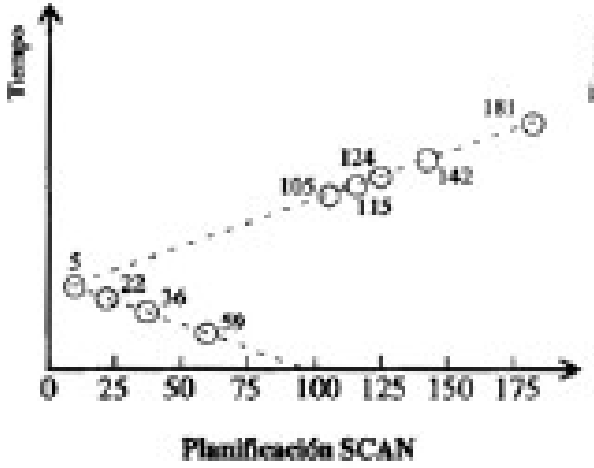
Tabla 2.20: Algoritmo SCAN de planificación del disco



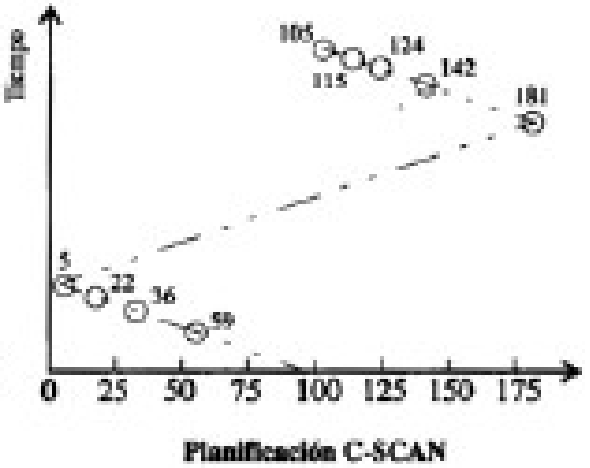
Planificación FCFS



Planificación SSTF



Planificación SCAN



Planificación C-SCAN