UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA

Microprocesadores para comunicaciones

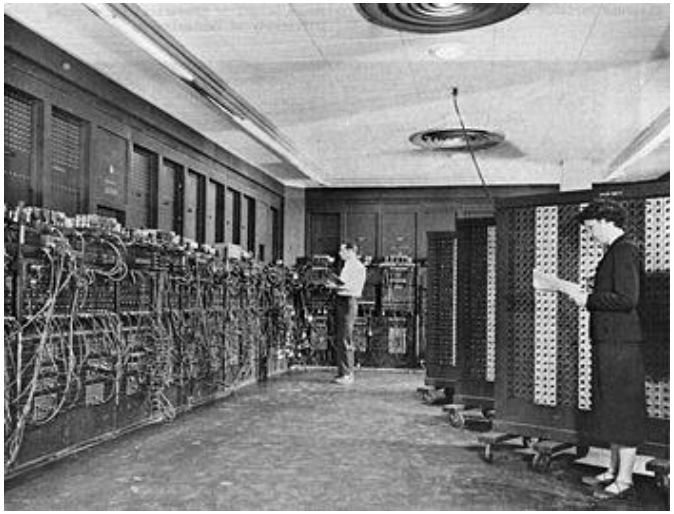
Escuela Técnica Superior de Ingenieros de Telecomunicación

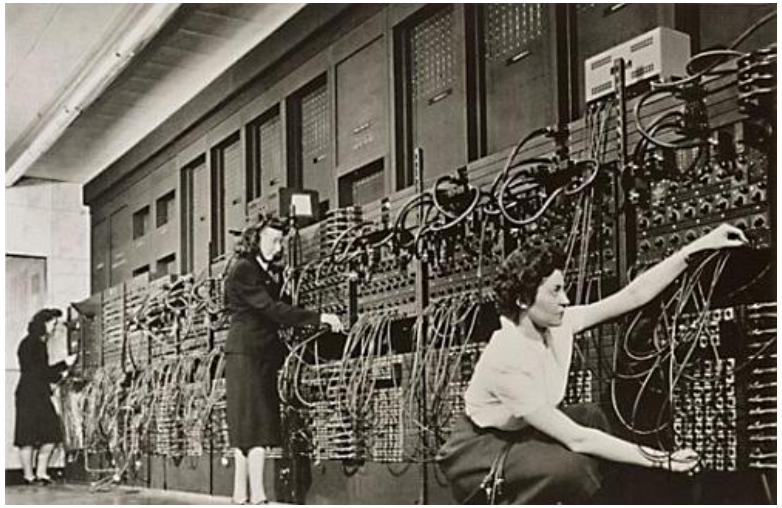
Organización y estructura de las memorias caché

Índice

Índice

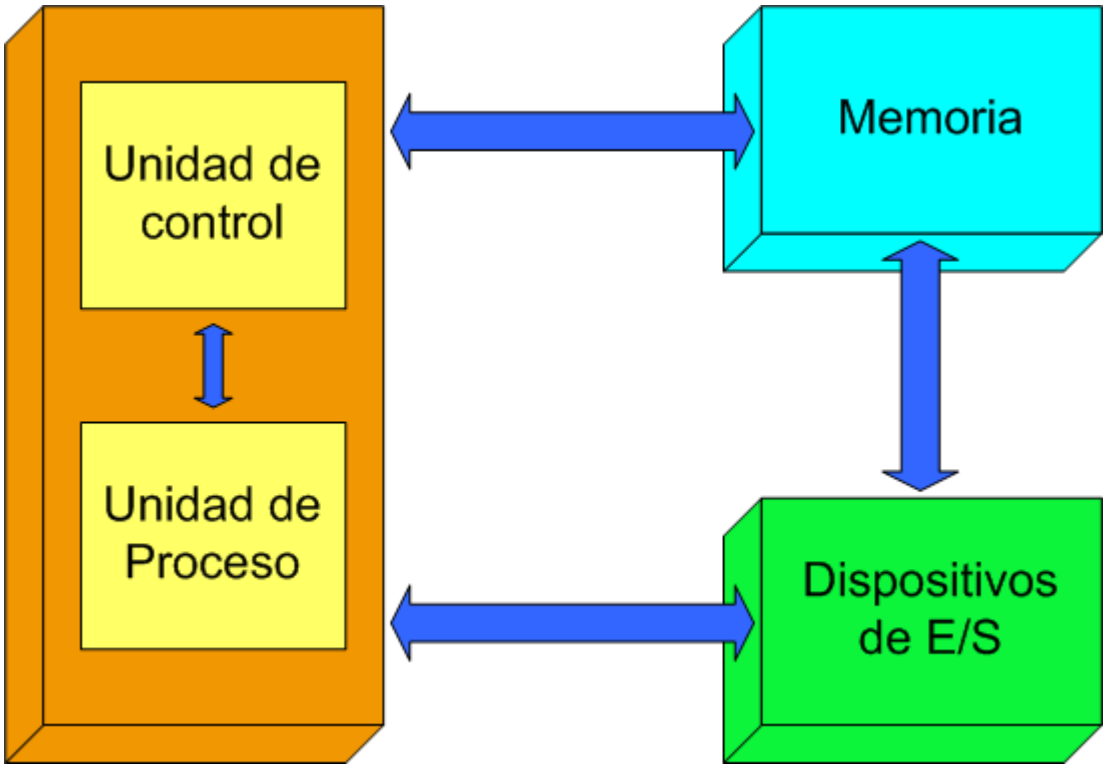
Introducción (I)





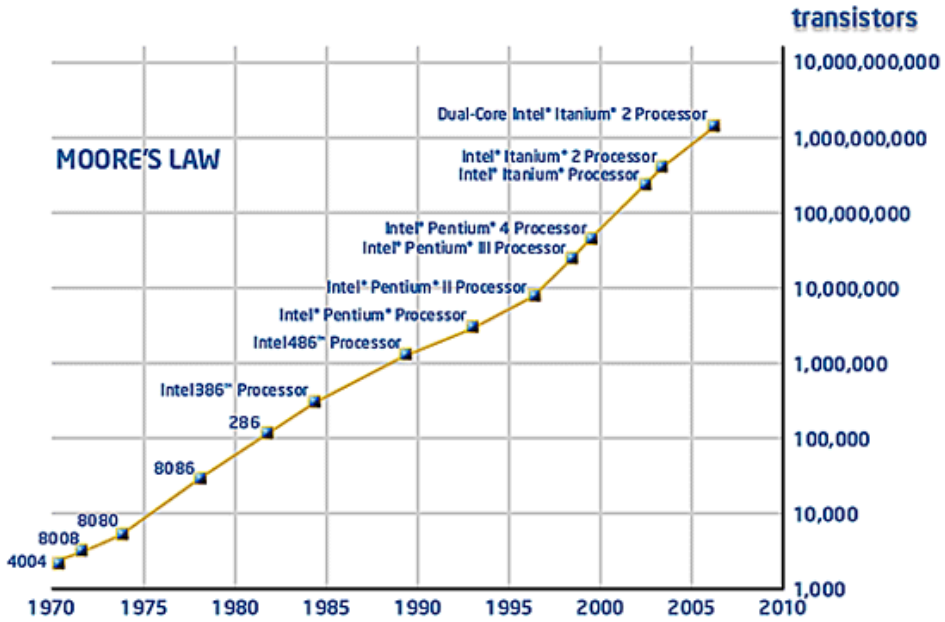
ENIAC (Electronic Numerical Integrator And Computer)

Introducción (II)



Arquitectura Von Neumann

Introducción (III)



Ley de Moore: el nivel de integración se duplicará cada año y medio

Introducción (IV)

- Las memorias no han experimentado el mismo grado de mejora en lo que se refiere al tiempo de acceso

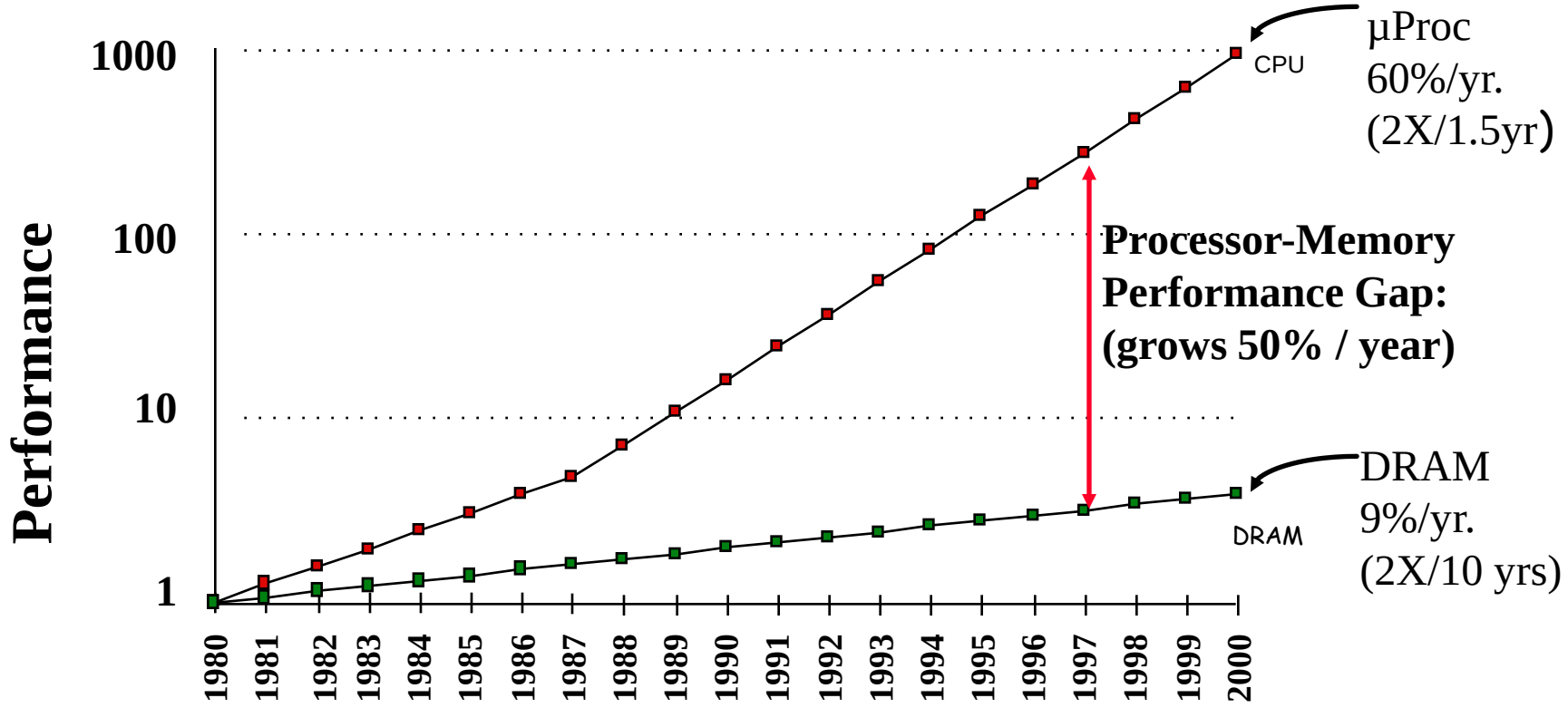
Evolución de la memoria *DRAM*

Año	Tamaño	Tiempo ciclo
1980	64 Kb	250 ns
1983	256 Kb	220 ns
1986	1 Mb	190 ns
1989	4 Mb	165 ns
1992	16 Mb	145 ns
1995	64 Mb	120 ns

1000:1!

2:1!

Introducción (V)



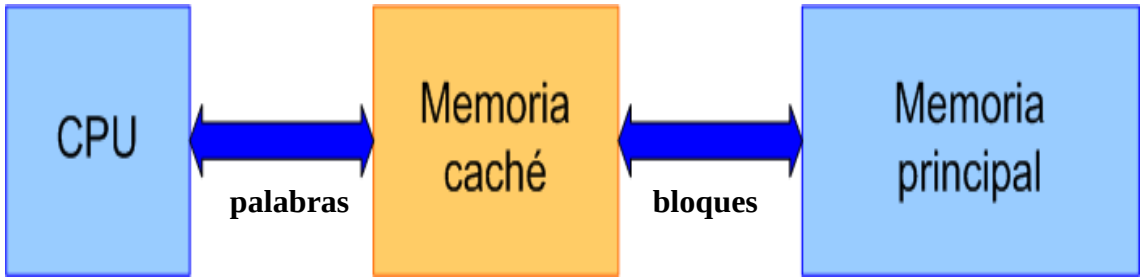
Diferencia de rendimiento entre memoria y microprocesador

Índice

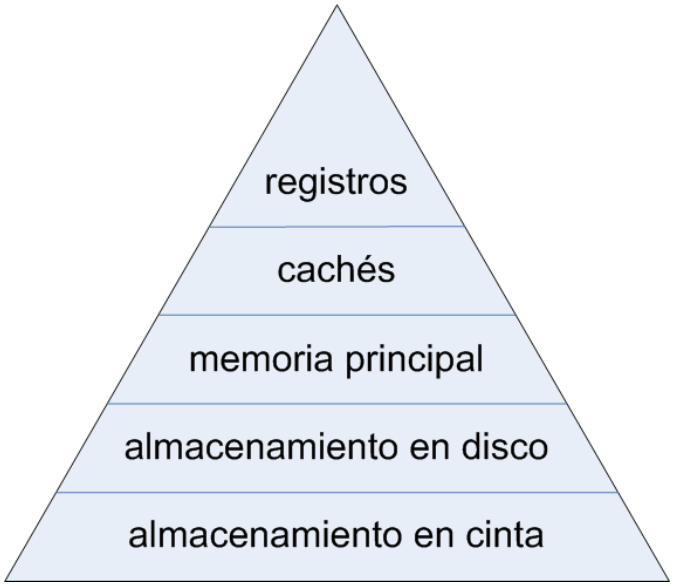
Niveles de jerarquía de memoria (I)

Memorias caché

- Memoria pequeña y rápida situada en el procesador y la memoria principal
- Almacena una copia de la información más recientemente utilizada
- Disminuye el tiempo de acceso a memoria
- Objetivo: Dar la impresión de que las referencias a memoria se realizan a una velocidad muy cercana a la del procesador

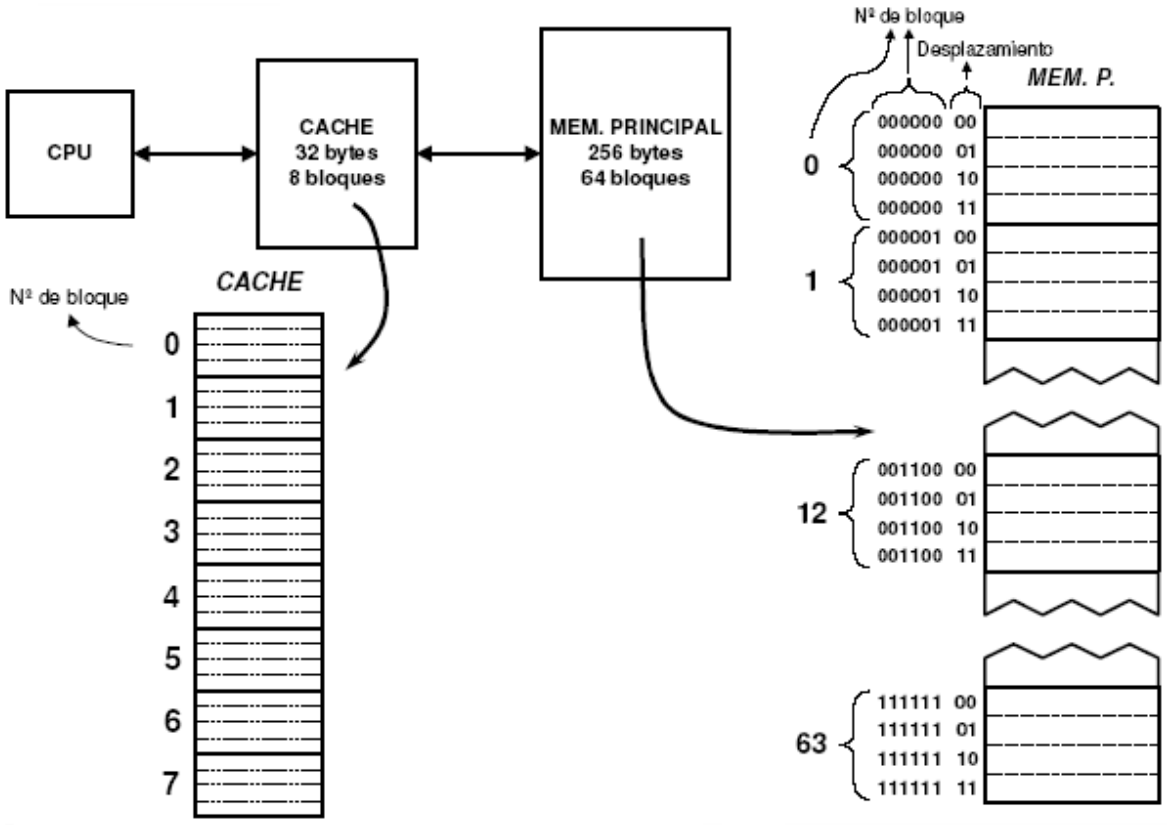


Niveles de jerarquía de memoria (II)



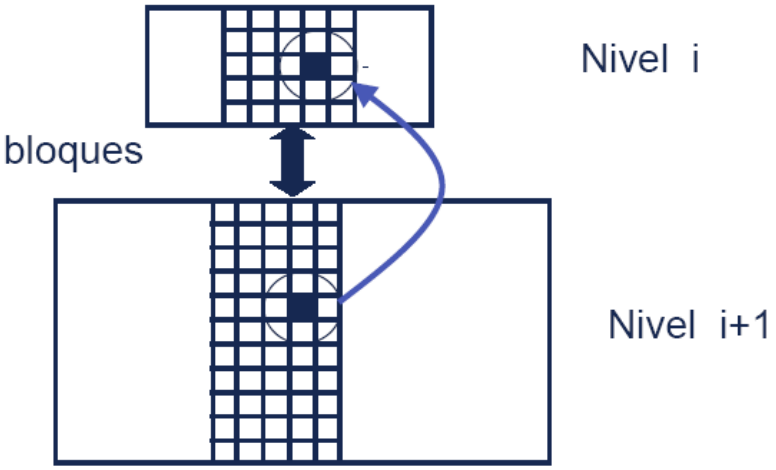
Velocidad	Tamaño	Coste
0,25 ns	500 B	Más alto
1 ns	64 KB	
100 ns	512 MB	
5 ms	100 GB	
		Más bajo

Niveles de jerarquía de memoria (III)



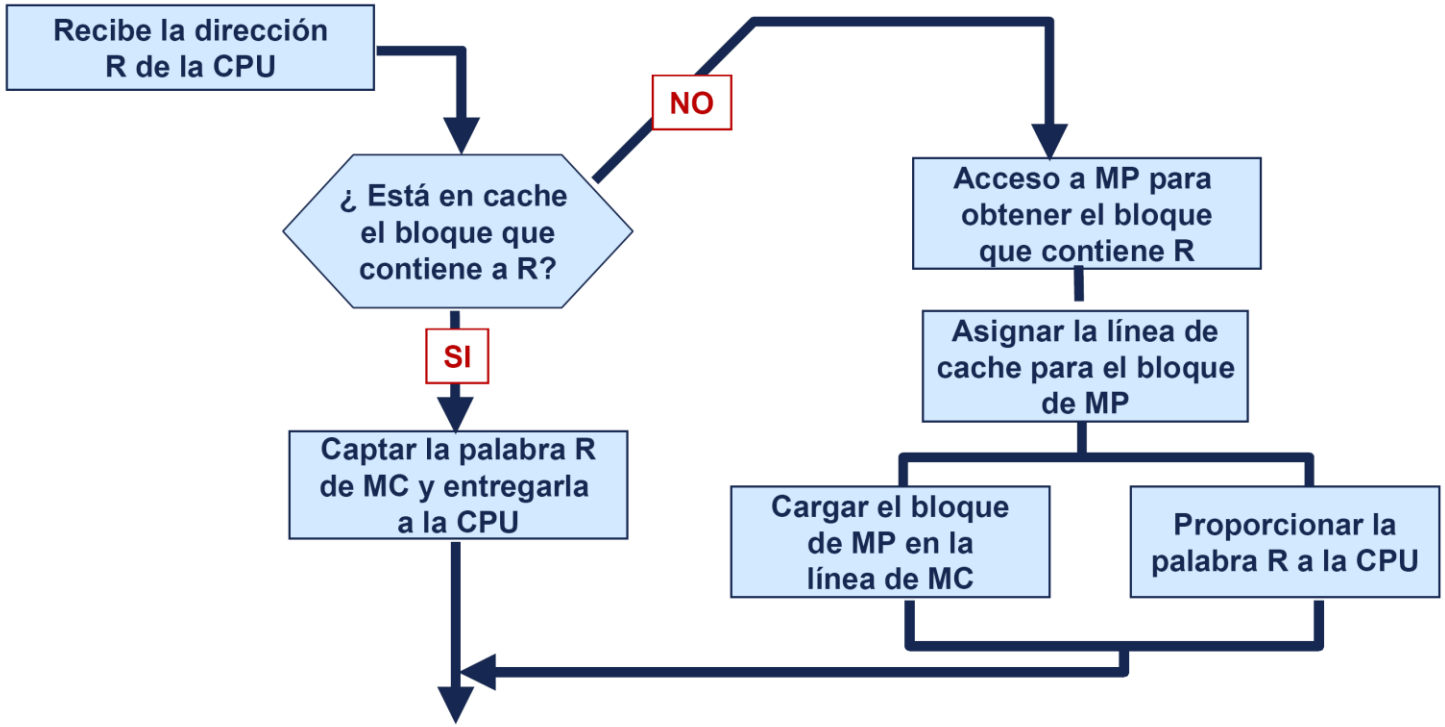
Niveles de jerarquía de memoria (IV)

- El sistema de memoria jerárquica debe hacer que en todo momento los datos que necesita la CPU se encuentren en el nivel más bajo de jerarquía
- La información fluye de un nivel superior a uno inferior a medida que se necesite



La unidad de transferencia entre la MP y la MCa es el bloque o línea

Niveles de jerarquía de memoria (V)



Operación de lectura de caché

Índice

Principio de localidad

- **Localidad Temporal:** Si se accede a una posición de memoria en un instante de tiempo determinado, existe una alta probabilidad de que se vuelva a acceder en los instantes siguientes.

- **Localidad Espacial:** Si se accede a una posición de memoria en un instante determinado, existe una alta probabilidad de que en los instantes siguientes se acceda a las posiciones de memoria cercanas.

Índice

Terminología

- **Acierto :** acceso en el que se encuentra el dato (*cache hit*)
- **Fallo**: acceso en el que no se encuentra el dato (*cache miss*)
- **Tasa de aciertos:** porcentaje de veces que se encuentra el dato que se busca en el nivel superior (*hit rate*)
- **Tiempo de acierto:** tiempo necesario para leer un dato de la memoria caché. Incluye determinar si fue *hit* o *miss*
- **Penalización por fallo:** tiempo adicional que se tarda en obtener un dato cuando se produce un fallo . Es el tiempo para reemplazar un bloque en el nivel superior con el correspondiente bloque del nivel inferior, más el tiempo para proveer ese bloque al procesador.

Índice

Políticas de ubicación

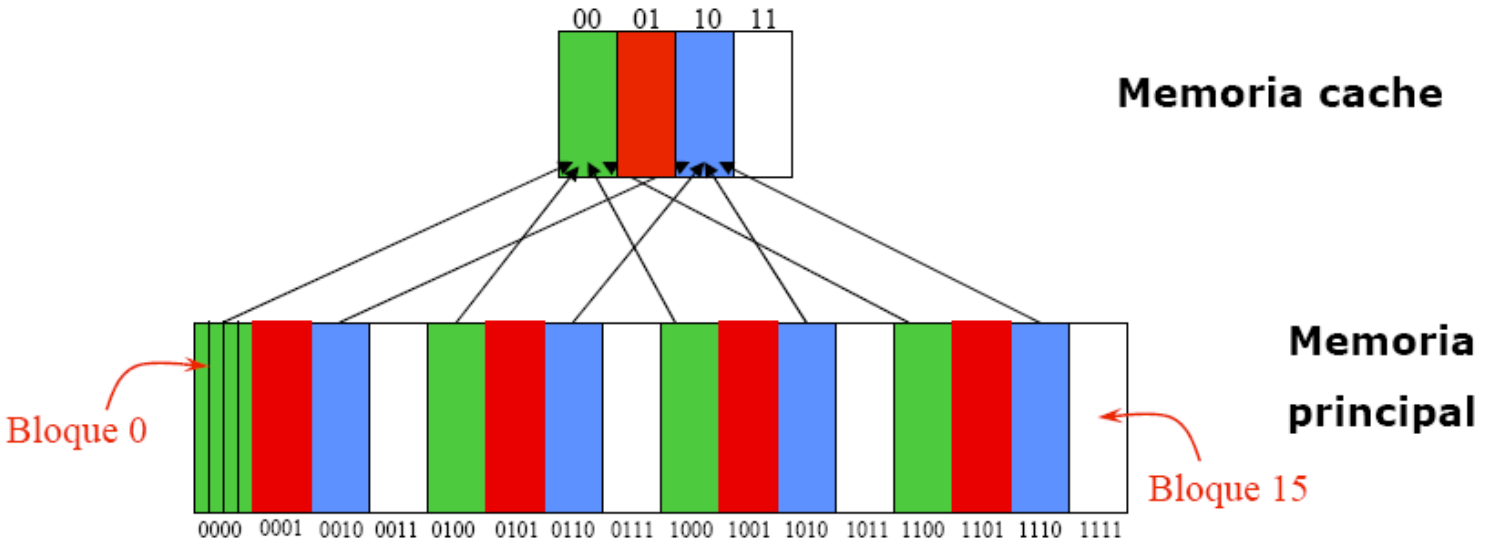
- Establece la correspondencia entre bloque de la MP y la MCa (dónde puede ubicarse un bloque en una caché)
- Cuando una caché solita un bloque al nivel inferior, tiene que decidir dónde lo ubica
- La posible ubicación de un bloque crea tres categorías en la organización de la cachés
  - Cachés de correspondencia directa (direct mapped cache)
  - Cachés completamente asociativa (fully associative cache)
  - Cachés de correspondencia asociativa por conjuntos (set associative cache)

Política de ubicación (I)

- Caché de correspondencia directa

Cada bloque de la memoria principal sólo puede ir en una línea de la caché

Número línea caché = número del bloque de la MP MODULO número de líneas en la caché



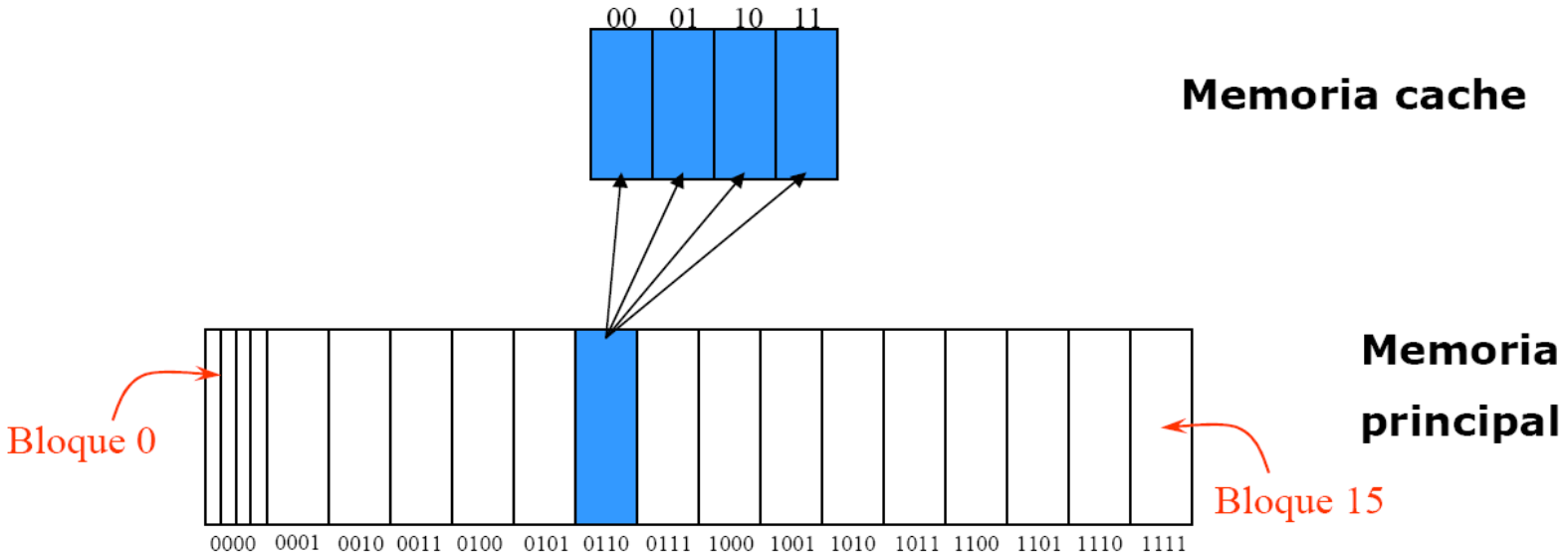
Política de ubicación (II)

- Caché de correspondencia directa
  - Ventajas:
    - Algoritmo de reemplazo trivial
  - Inconvenientes:
    - Incremento de la tasa de fallos de caché, si dos bloques de memoria principal, que corresponden a un mismo bloque de la caché, se utilizan de forma alternativa

Política de ubicación (III)

- Caché de correspondencia totalmente asociativa

  Cada bloque de la memoria principal puede ir en cualquier posición de la caché



Política de ubicación (IV)

- Caché de correspondencia totalmente asociativa

  - Ventajas:

    - Flexibilidad ya que permite la implantación de un gran variedad de algoritmos de reemplazo

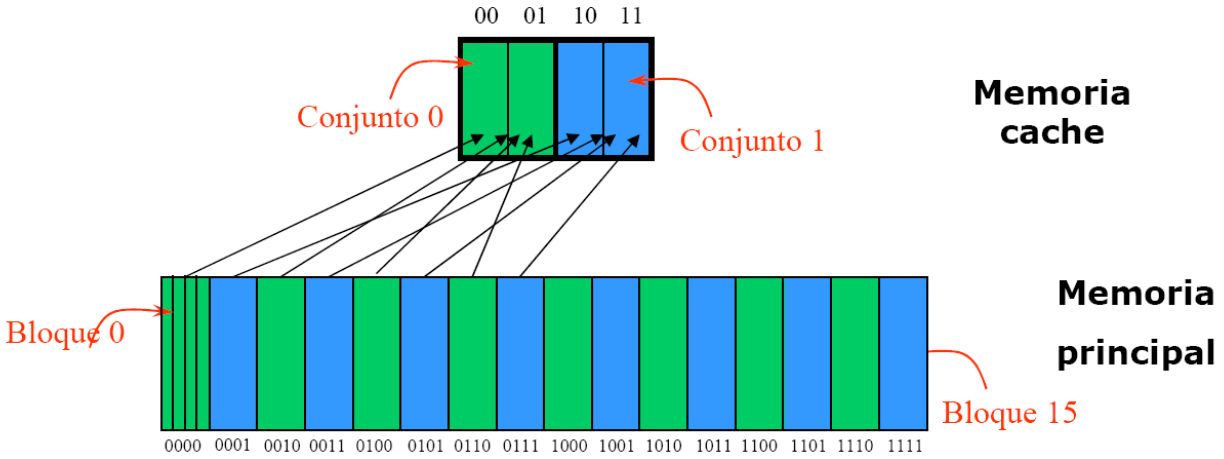
  - Inconvenientes:

    - Coste de las comparaciones

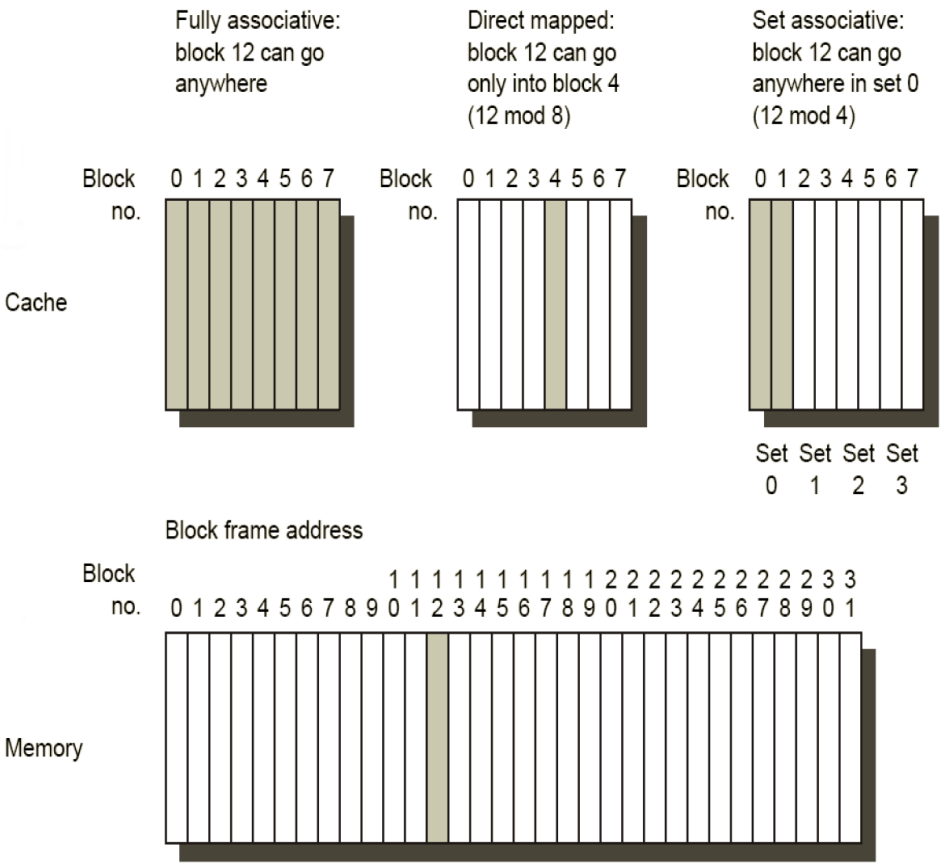
Política de ubicación (V)

- Caché de correspondencia asociativa por conjuntos
  - La memoria caché se divide en varios conjuntos de bloques
  - Los bloques de MP se asignan mediante correspondencia directa a los conjuntos de la caché, pero dentro de éstos el bloque de ubicación se elige libremente

Nº de conjunto = Nº bloque MP MODULO **Nº conjuntos caché**



Política de ubicación (VI)



Índice

Arquitecturas hardware (I)

❖ Correspondencia directa

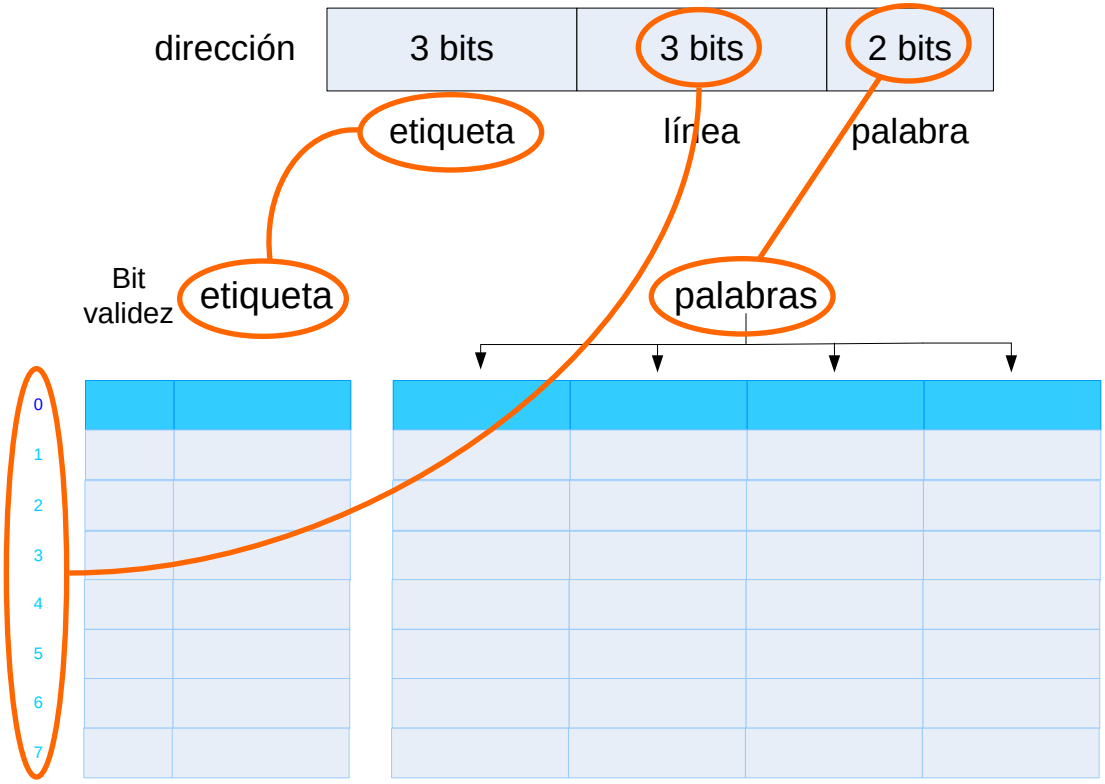
Cada línea de la caché está compuesta por tres campos:

- Datos
- Etiqueta: Identifica el bloque de MP
- Bit de validez

Bit validez	Etiqueta	Datos

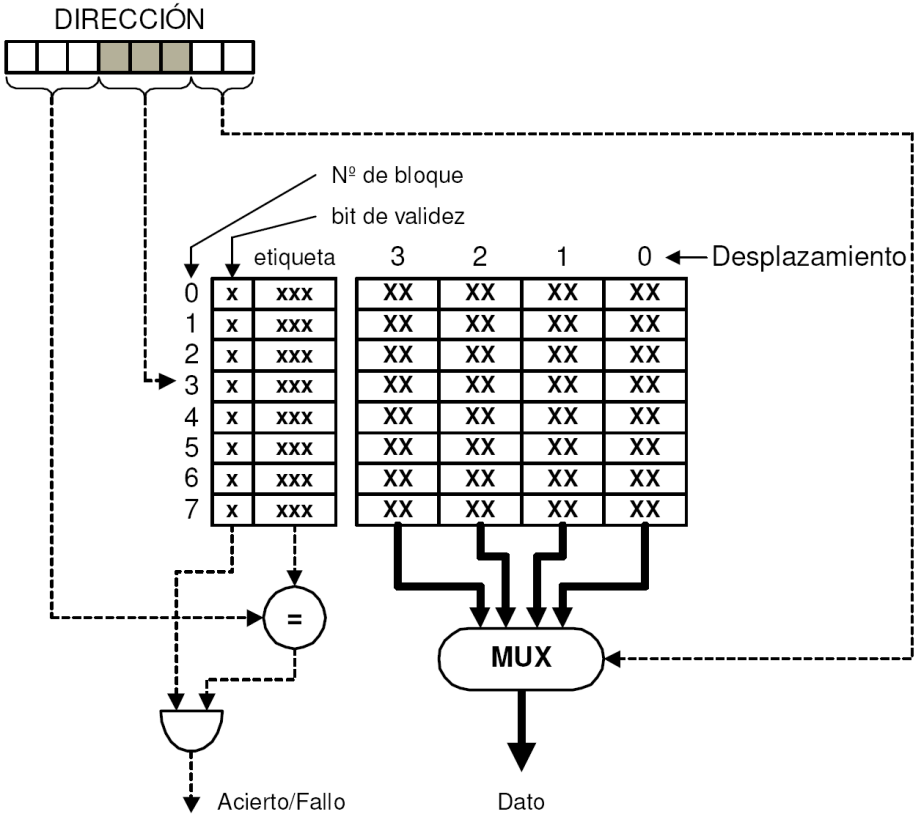
Arquitecturas hardware (II)

❖ Correspondencia directa



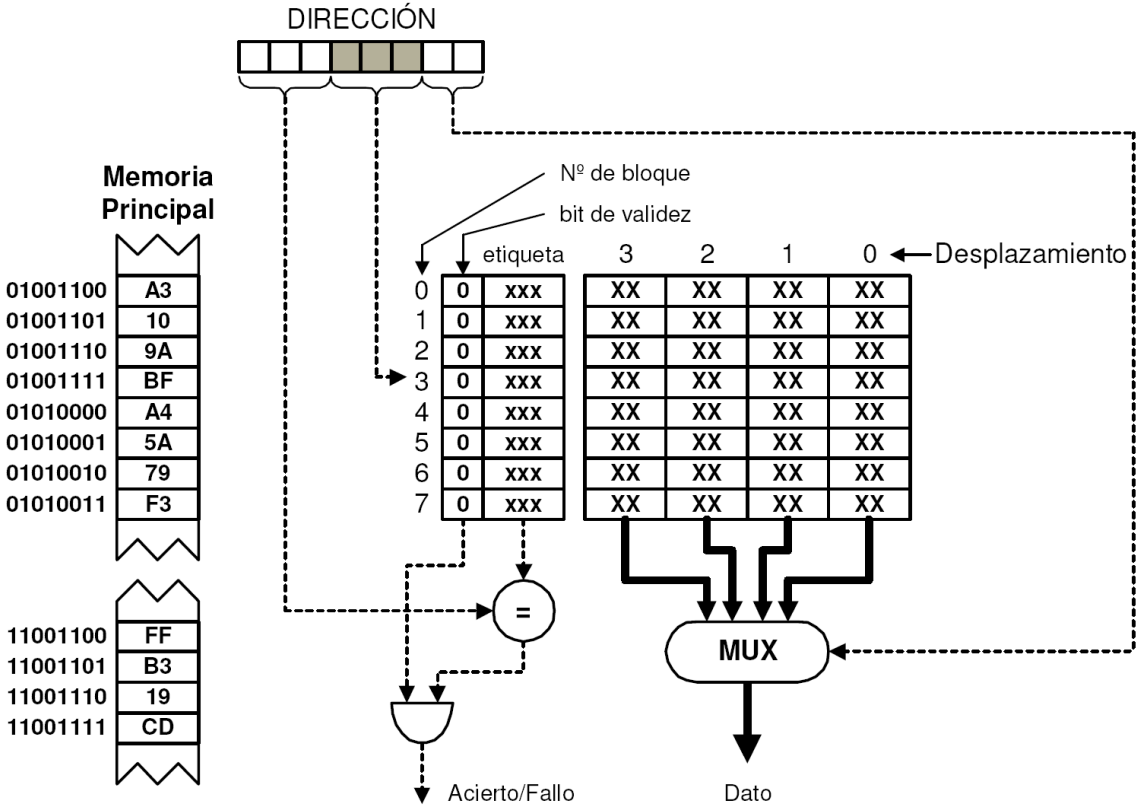
Arquitecturas hardware (III)

❖ Correspondencia directa



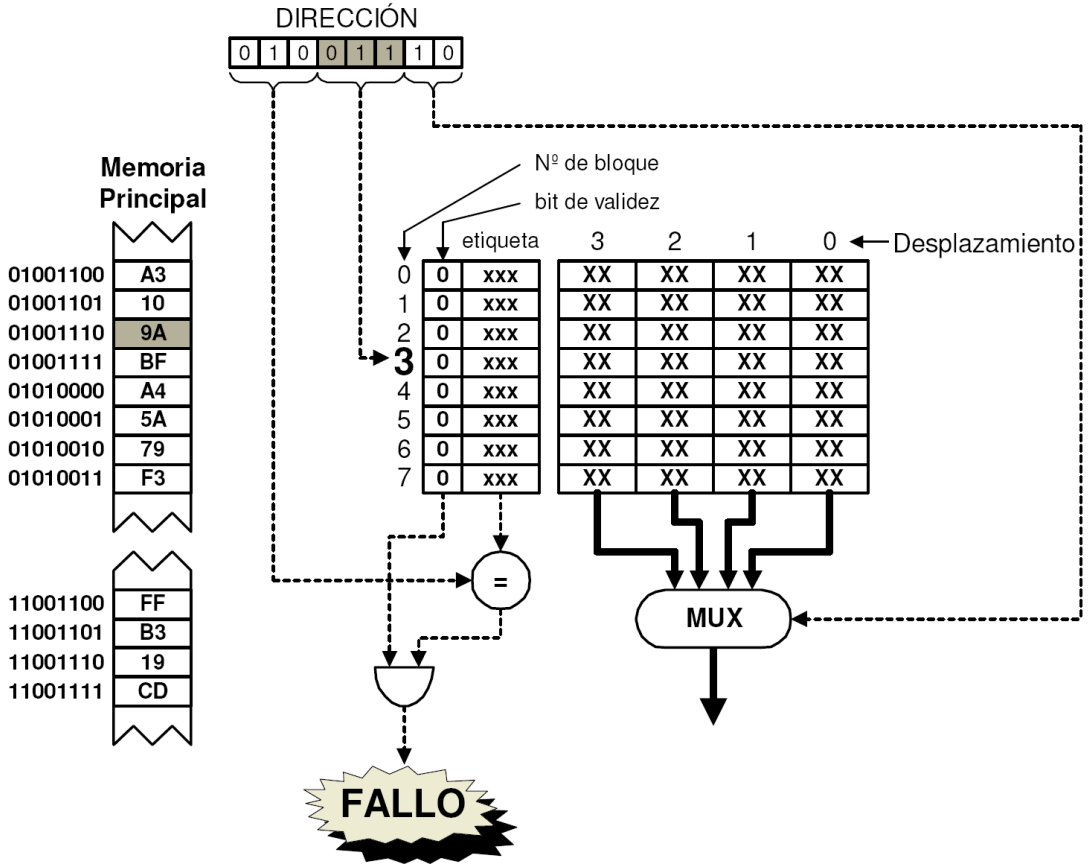
Funcionamiento cache de correspondencia directa (I)

- *Situación inicial: cache vacía*



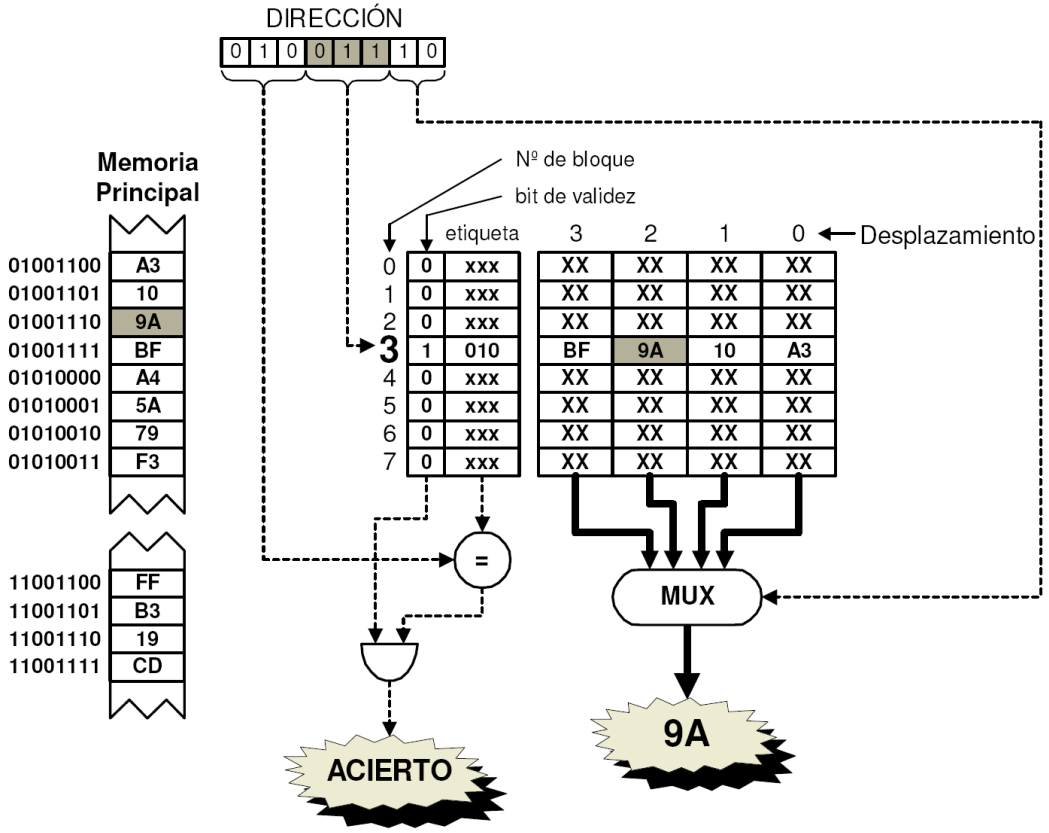
Funcionamiento cache de correspondencia directa (II)

- *Petición de lectura sobre la dirección 01001110b (4Eh): Fallo*



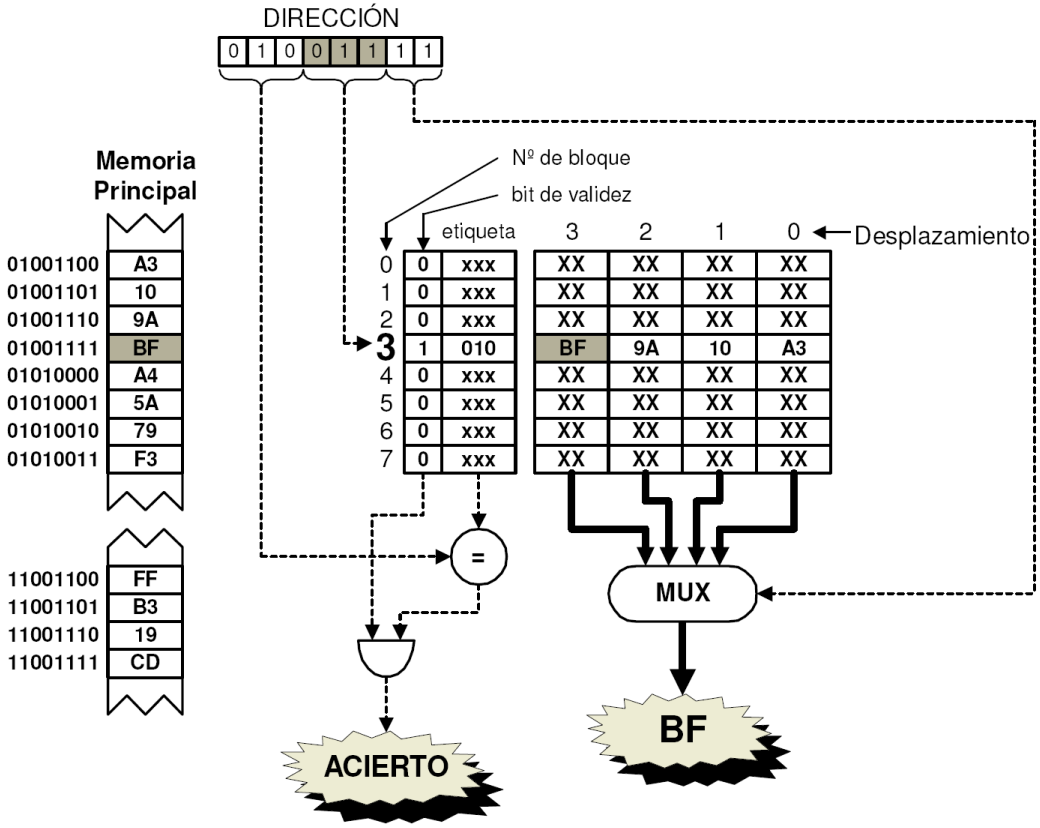
Funcionamiento cache de correspondencia directa (III)

- *Actualización del bloque 3*
- *La cache sirve el dato pedido*



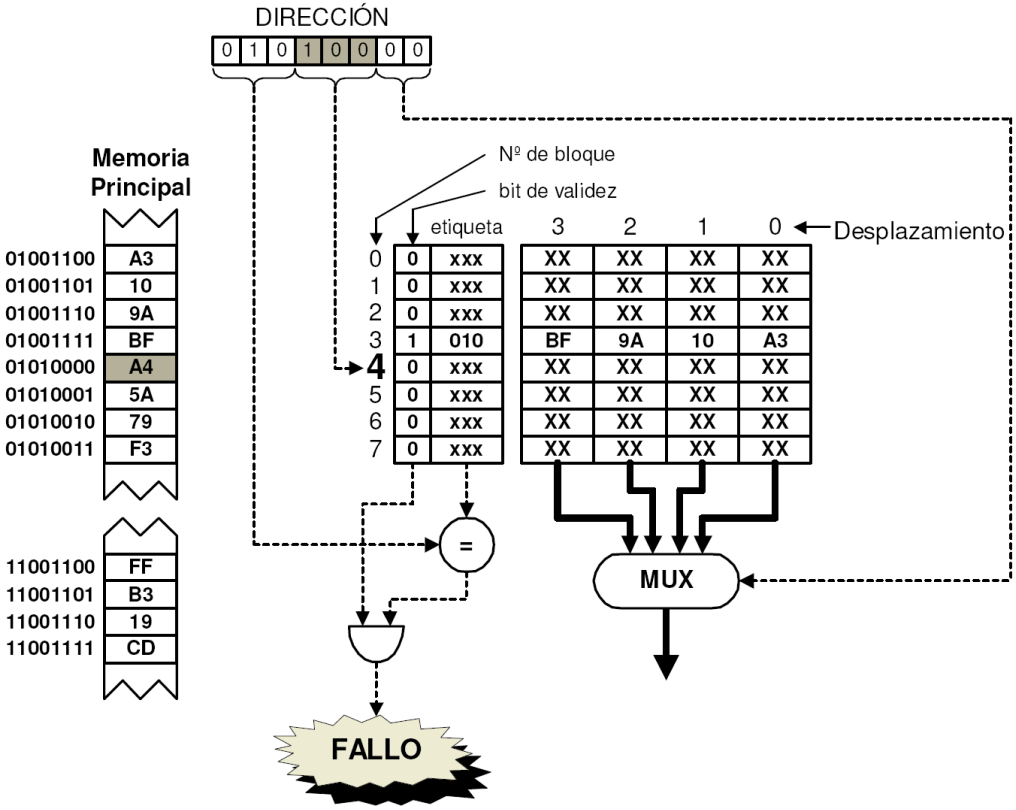
Funcionamiento cache de correspondencia directa (IV)

- *Petición de lectura sobre la dirección 01001111b (4Fh): Acierto*
- *La cache sirve el dato pedido*



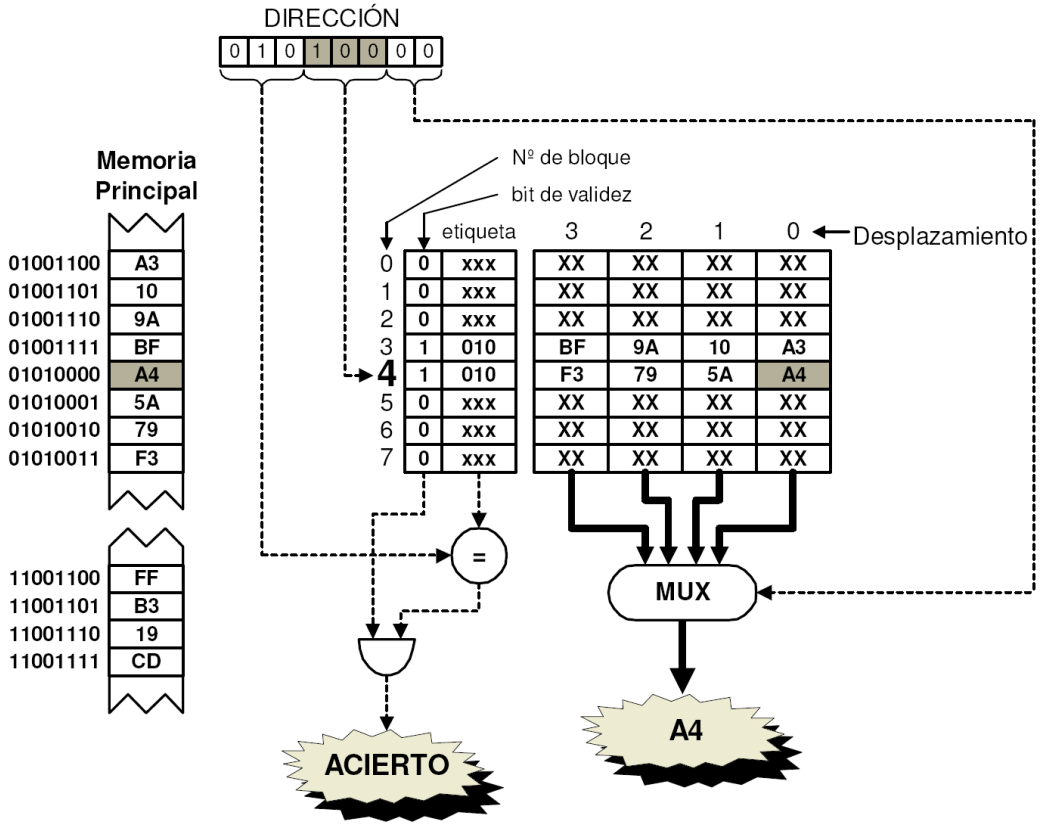
Funcionamiento cache de correspondencia directa (V)

- *Petición de lectura sobre la dirección 01010000b (50h): Fallo*



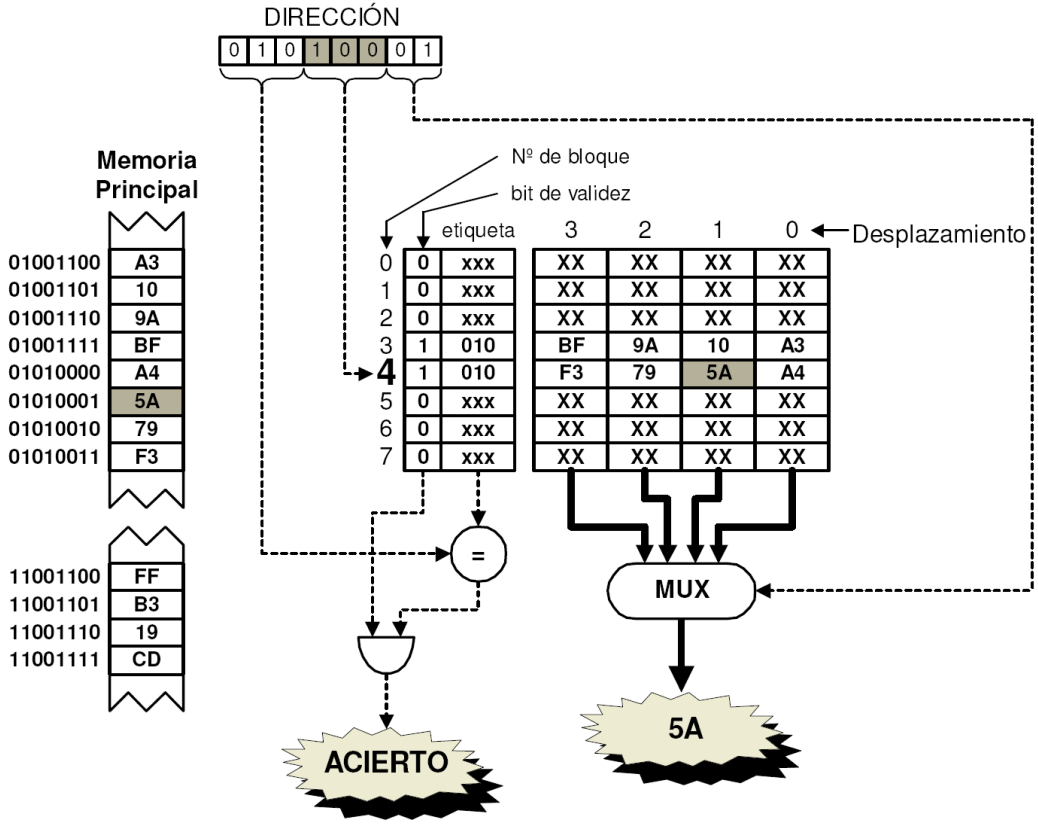
Funcionamiento cache de correspondencia directa (VI)

- *Actualización del bloque 4*
- *La cache sirve el dato pedido*



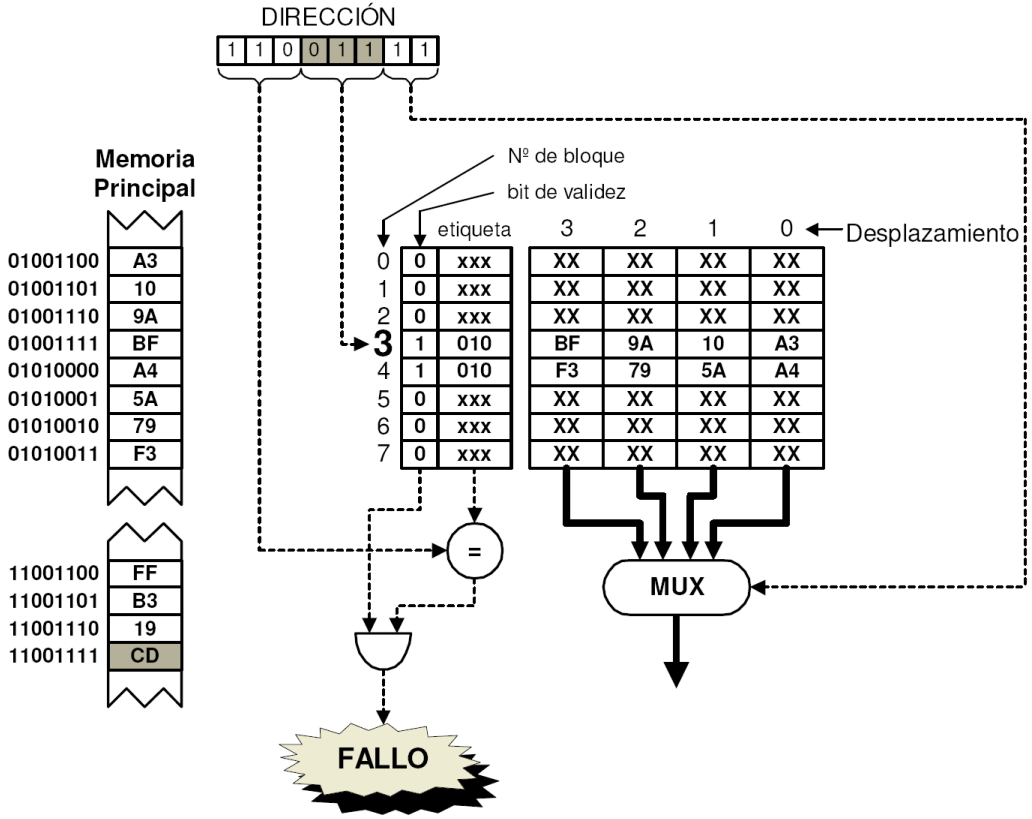
Funcionamiento cache de correspondencia directa (VII)

- *Petición de lectura sobre la dirección 01010001b (51h): Acierto*
- *La cache sirve el dato pedido*



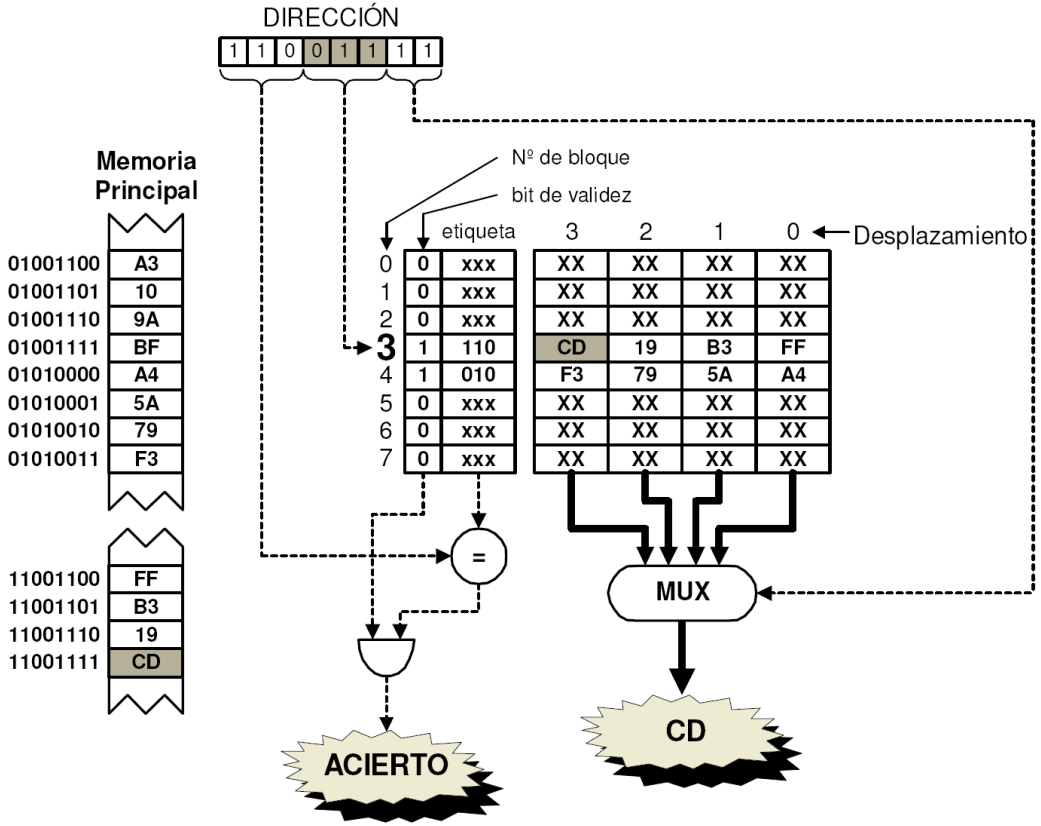
Funcionamiento cache de correspondencia directa (VIII)

- *Petición de lectura sobre la dirección 11001111b (CFh): Fallo*



Funcionamiento cache de correspondencia directa (IX)

- *Reemplazo del bloque 3*
- *La cache sirve el dato pedido*



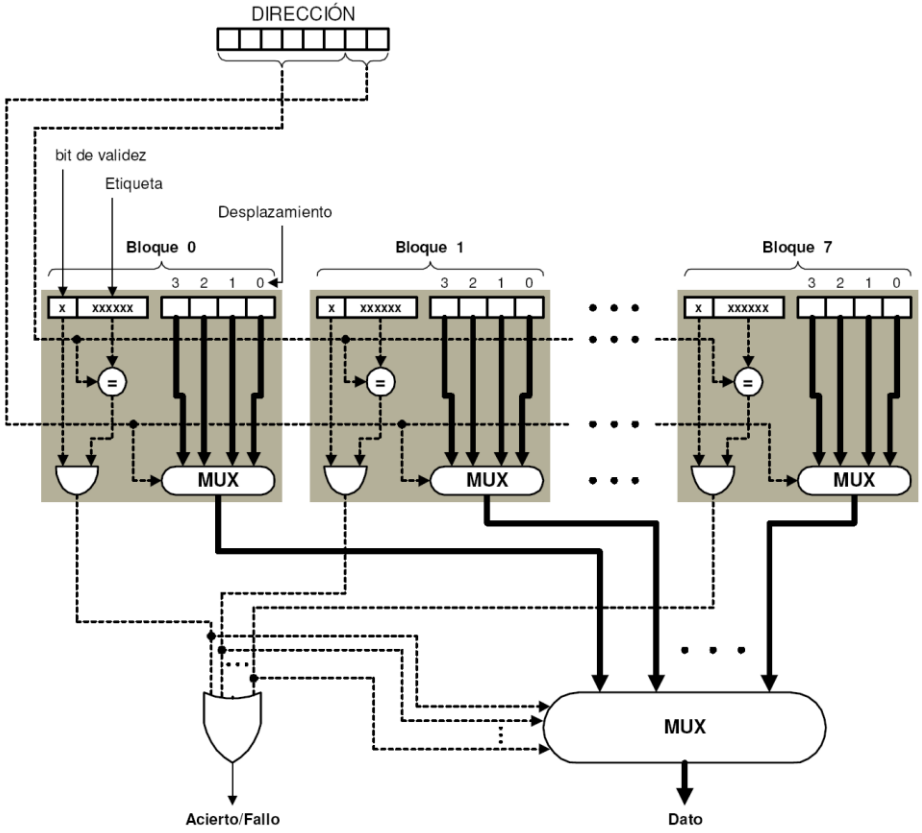
Funcionamiento cache de correspondencia directa (X)

❖ Ejemplo: Dividir la dirección de la memoria principal en campos para una política de ubicación de correspondencia directa

- Memoria principal de 1Mbytes
- Memoria caché de 1Kbytes
- Tamaño del bloque de 8 bytes

Arquitecturas hardware (IV)

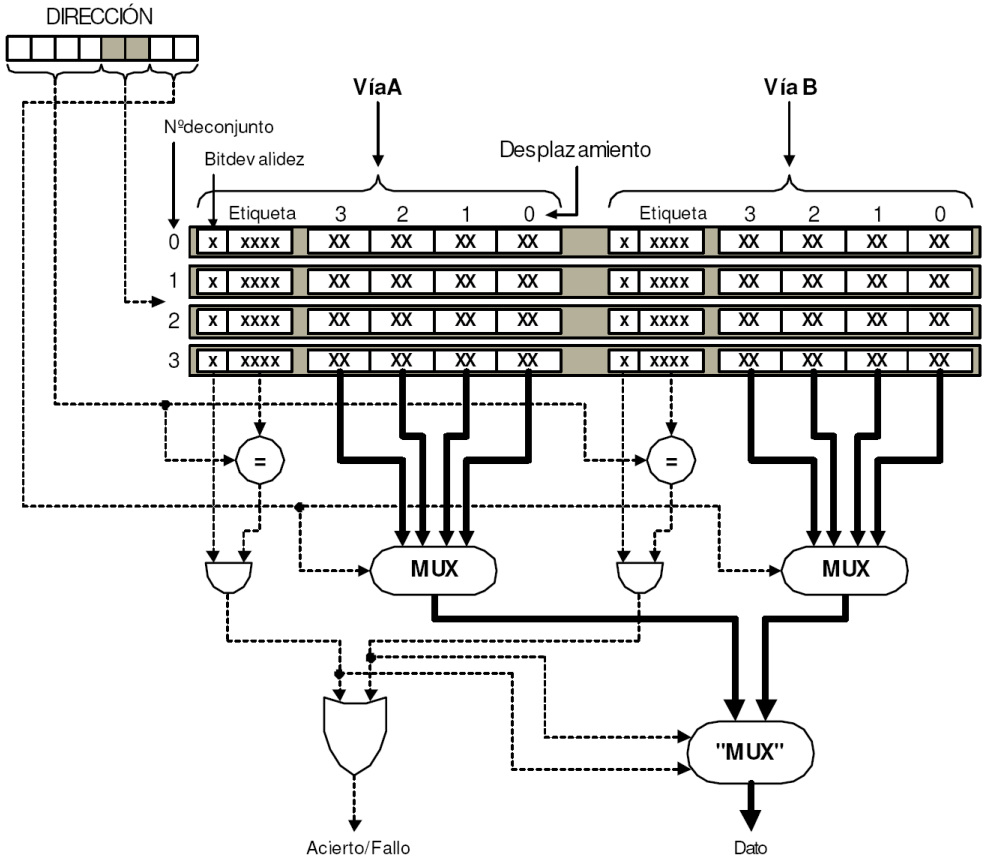
❖ Estructura de una caché totalmente asociativa



Arquitecturas hardware (V)

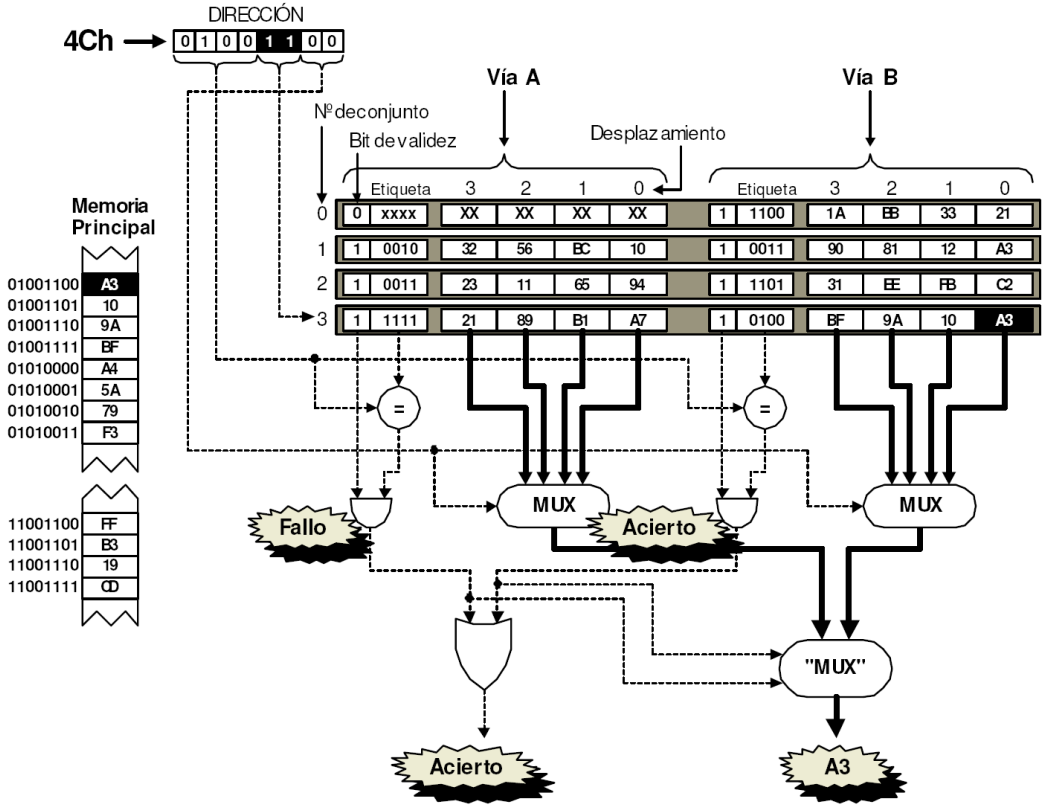
❖ Estructura de una caché asociativa por conjuntos

- *Estructura: 4 conjuntos de 2 bloques*



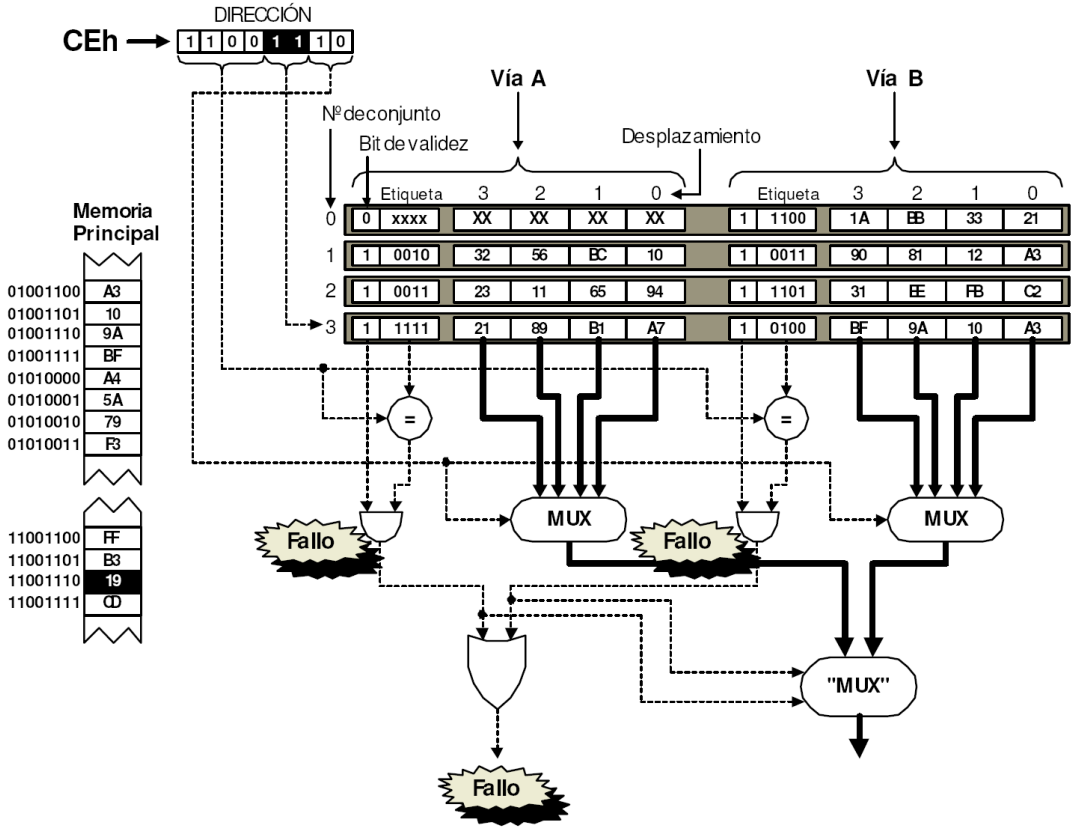
caché asociativa por conjuntos (I)

- *Petición de lectura sobre la dirección 01001100b (4Ch): ACIERTO*
- *La cache sirve el dato solicitado*



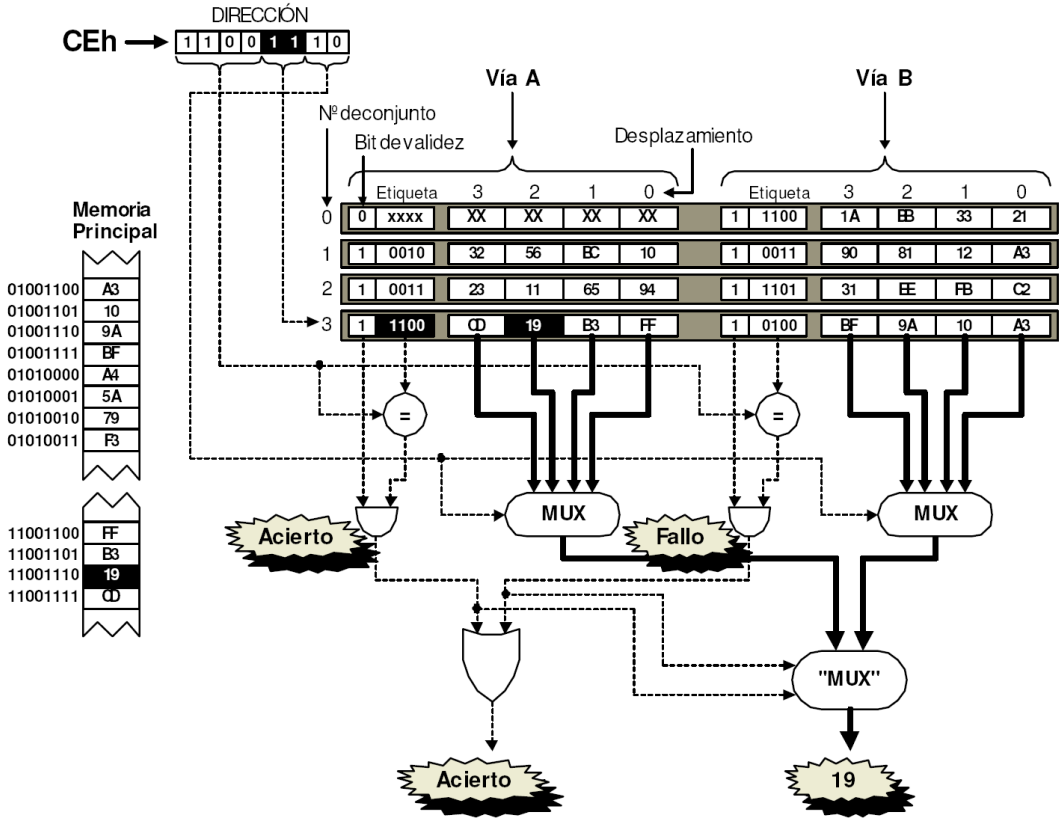
caché asociativa por conjuntos (II)

- *Petición de lectura sobre la dirección 11001110b (CEh) : FALLO*



caché asociativa por conjuntos (III)

- *Actualización del bloque ubicado en la vía A del conjunto 3*
- *La* cache *sirve el dato solicitado*



Caché asociativa por conjuntos (IV)

- Ejemplo de correspondencia entre dirección de memoria principal y una memoria caché asociativa por conjuntos

  - Memoria principal de 1Mbytes
  - Memoria caché de 1Kbytes
  - Tamaño del bloque de 8 bytes
  - Tamaño del conjunto 2 bloques

Índice

Políticas de reemplazo

- Determina cuándo y qué bloque de memoria caché debe abandonarla cuando no existe espacio disponible para un bloque entrante.

  - Aleatoria: Se escoge una línea del espacio de reemplazamiento al azar.
  - FIFO: Consiste en reemplazar la línea que ha permanecido en la MCa el mayor periodo de tiempo
  - Menos recientemente usado (LRU Least-recently used): Se sustituye aquella línea de MCa que hace más tiempo que no se ha utilizado.
  - LFU (Least Frequently Used): Se sustituye la línea del espacio de reemplazamiento que haya sido menos referenciada.

Tasa de fallos vs política de reemplazo

Tasa de fallos						
Tamaño	2 Vías		4 Vías		8 Vías	
	LRU	Aleatorio	LRU	Aleatorio	LRU	Aleatorio
16 KB	5,18 %	5,69 %	4,67 %	5,29 %	4,39 %	4,96 %
64 KB	1,88 %	2,01 %	1,54 %	1,66 %	1,39 %	1,53 %
256 KB	1,15 %	1,17 %	1,13 %	1,13 %	1,12 %	1,12 %

Índice

Políticas de actualización (I)

- Determina cuándo se actualiza la información en MP al haberse producido una escritura en MCa para evitar problemas de falta de coherencia
  - **Escritura inmediata (Write Through):** Cuando se escribe un bloque en memoria caché se actualiza directamente la información también en memoria principal
    - Ventajas la realización es muy sencilla y asegura la coherencia
    - Inconvenientes produce mucho tráfico entre memoria y el procesador debe esperar que se complete la escritura lo que lleva al empleo de buffer de escritura
  - **Escritura aplazada (Write back):** Consiste en escribir en MCa y únicamente se escribe en MP si el bloque a reemplazar ha sido modificado
    - Ventajas produce menos tráfico entre la memoria y el procesador y las escrituras se hacen a la velocidad de la caché
    - Inconvenientes el diseño es más complejo ya que hacen falta bits extras de control.

Políticas de actualización (II)

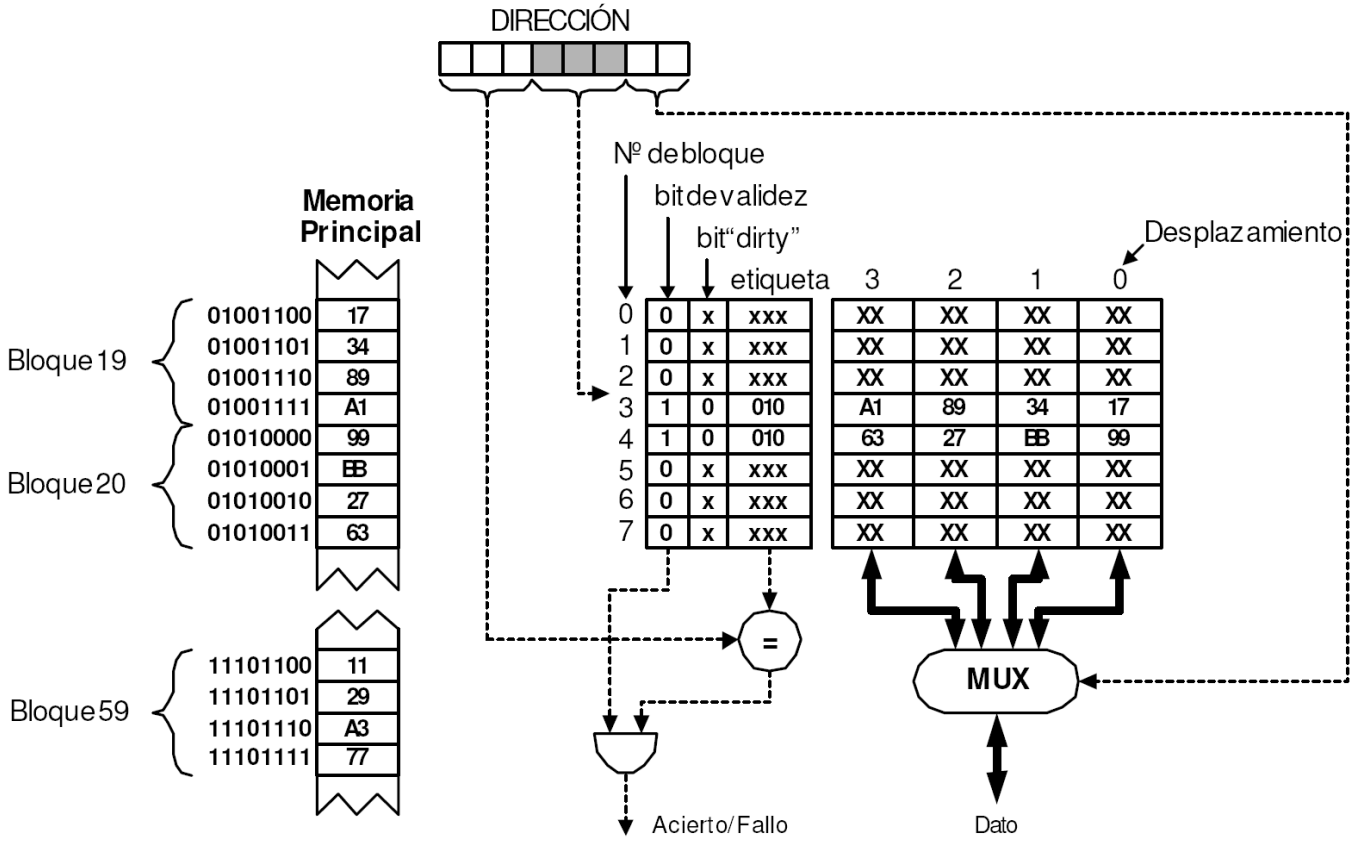
- Existen dos formas de actuar en el caso en que un acceso de escritura produzca un fallo:

  - **Escritura con ubicación (Write with allocate):** se suele asociar con escritura aplazada. Consiste en llevar el bloque que produce el fallo de MP a MCa y a continuación realizar la escritura en MCa

  - **Escritura sin ubicación (Write with no allocate):** se suele asociar con escritura inmediata. Consiste en realizar únicamente la escritura sobre la MP cuando se produce un fallo

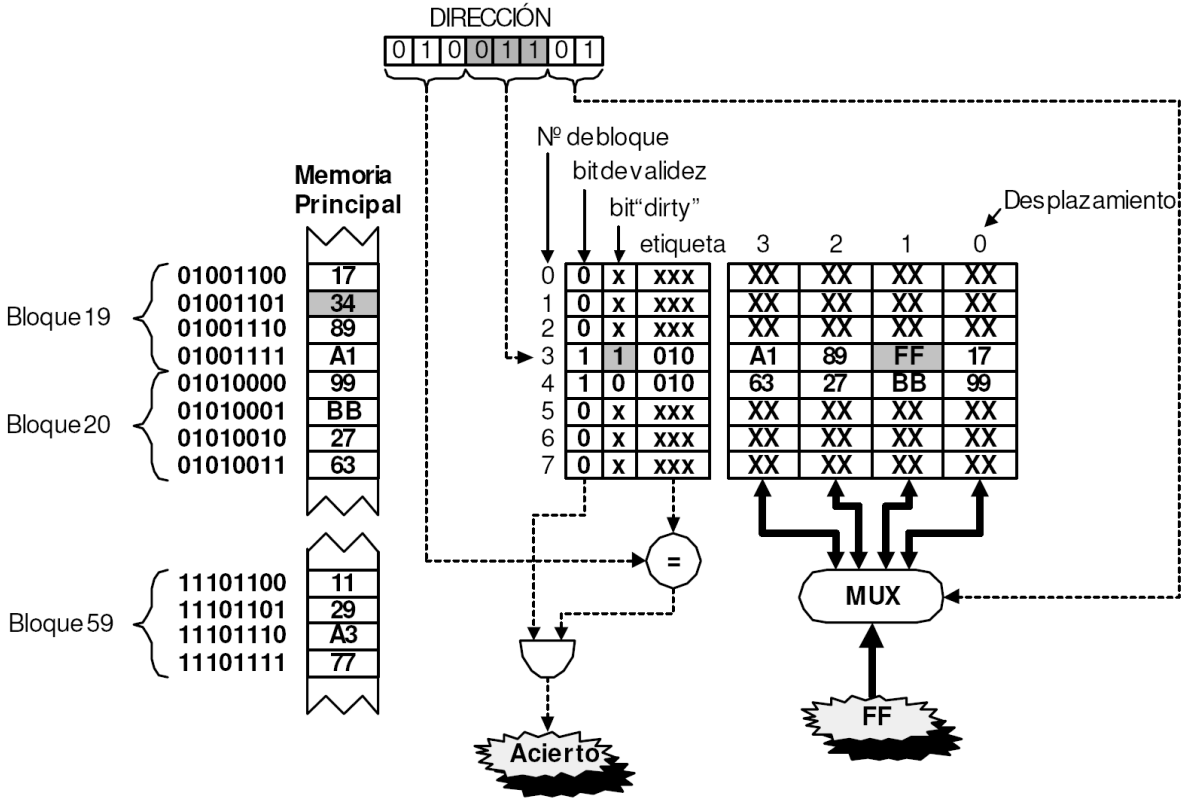
Funcionamiento de la caché con write-back (I)

- *Situación inicial de la cache y la M. P.*



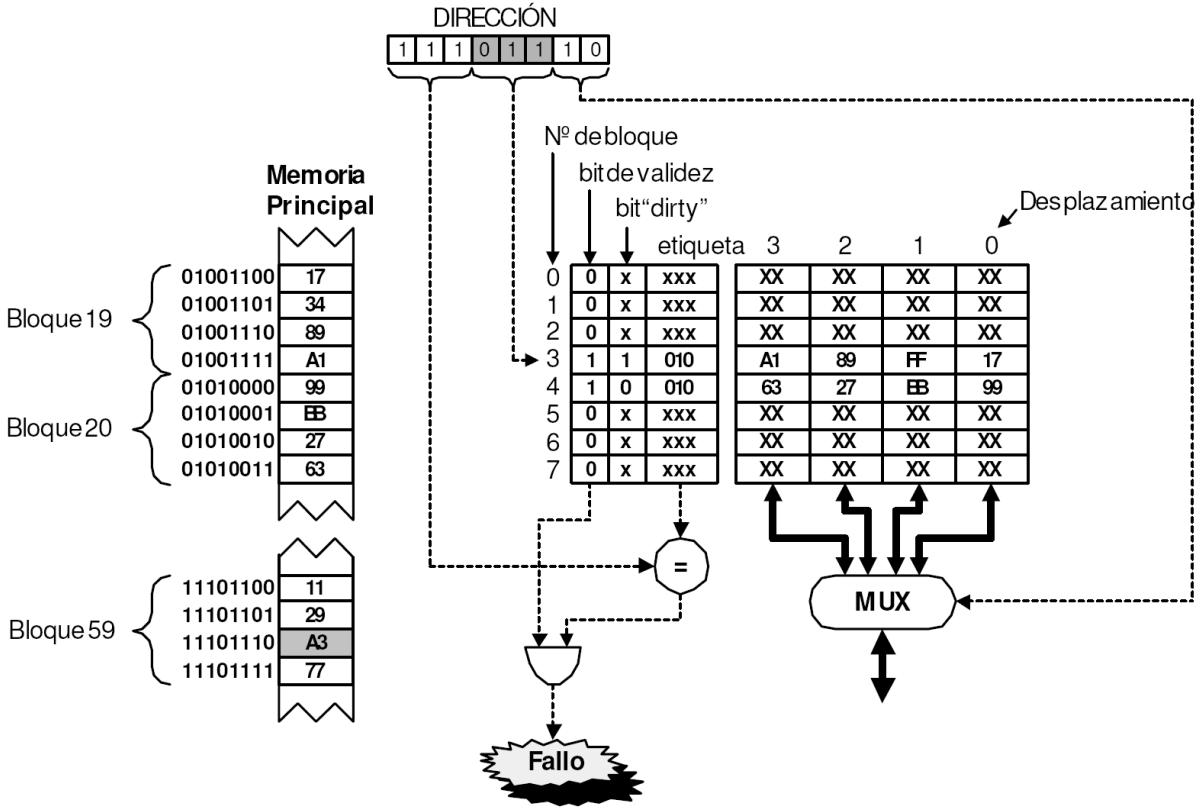
Funcionamiento de la caché con write-back (II)

- *Petición de escritura sobre la dirección 01001101b (bloque 19): Acierto*
- *Se escribe el dato FFh en la cache (bloque 3)*
- *El bloque correspondiente (bloque 3) se marca como "sucio"*



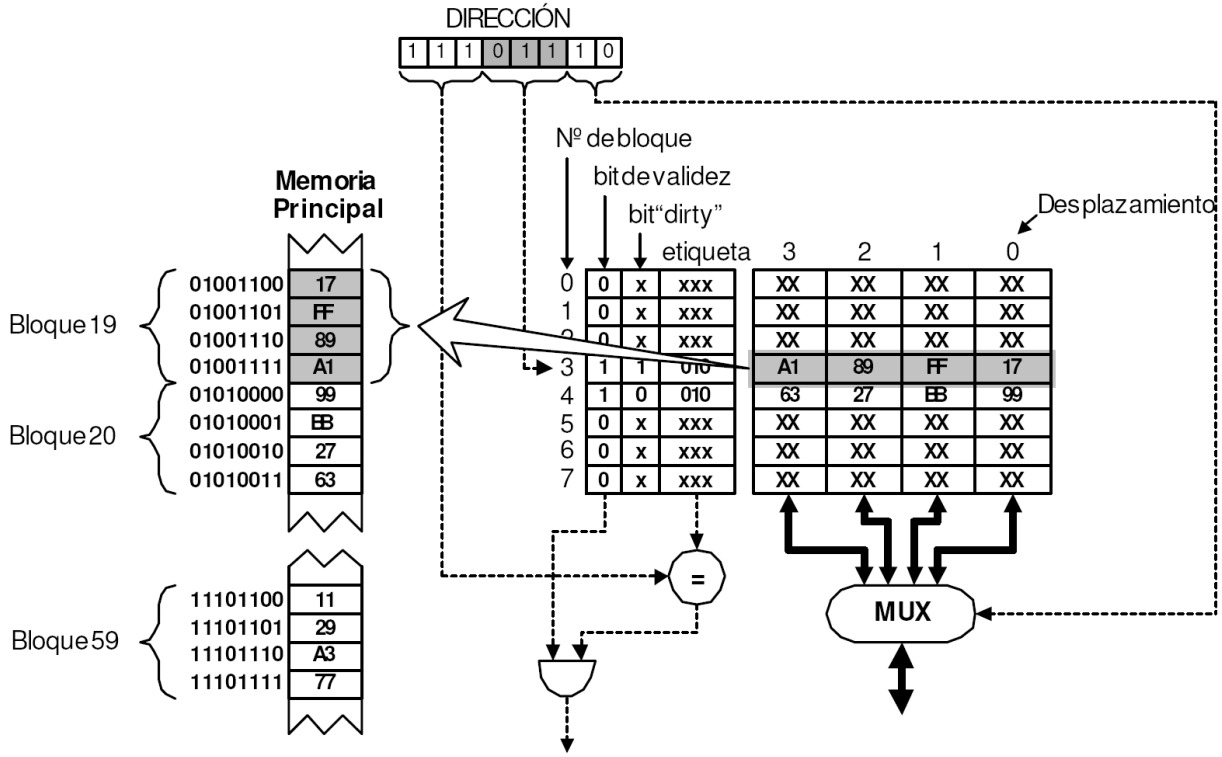
Funcionamiento de la cache con write-back (III)

- *Petición de lectura sobre la dirección 11101110b (bloque 59): Fallo*



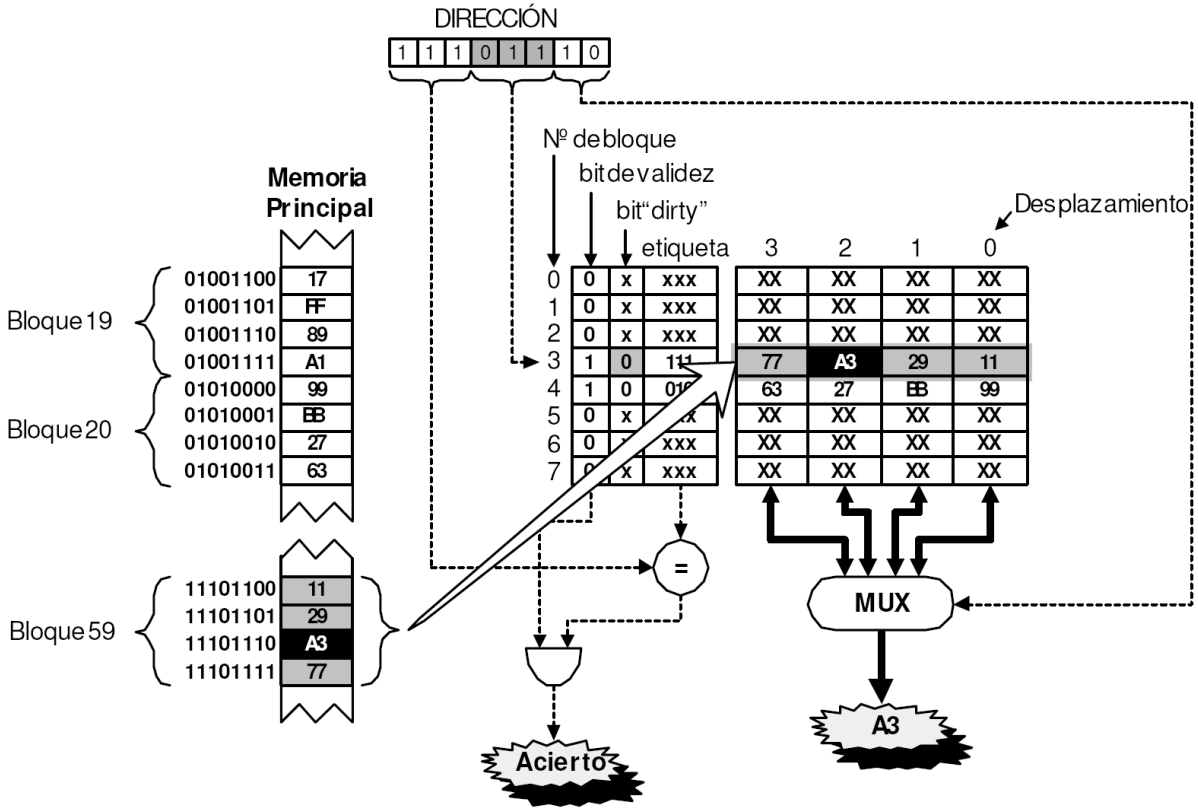
Funcionamiento de la cache con write-back (IV)

- *Actualización del bloque 19 de la M. P. (se salvaguarda así el bloque "sucio", que es el que contiene la información más actual)*



Funcionamiento de la cache con write-back (V)

- *Reemplazo del bloque 3 de la cache*
- *La cache sirve el dato pedido*



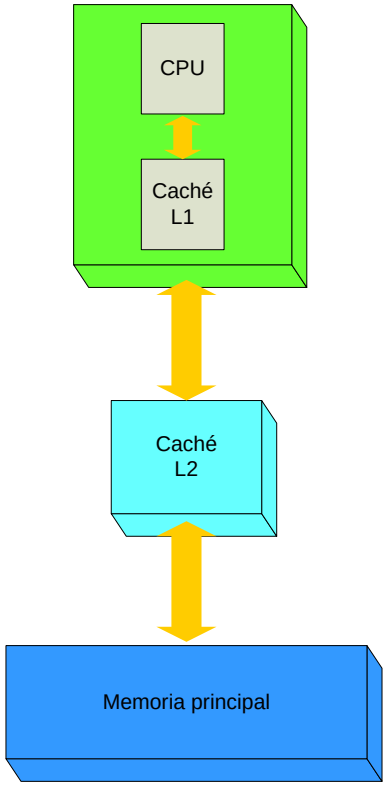
Índice

Memorias cache multinivel (I)

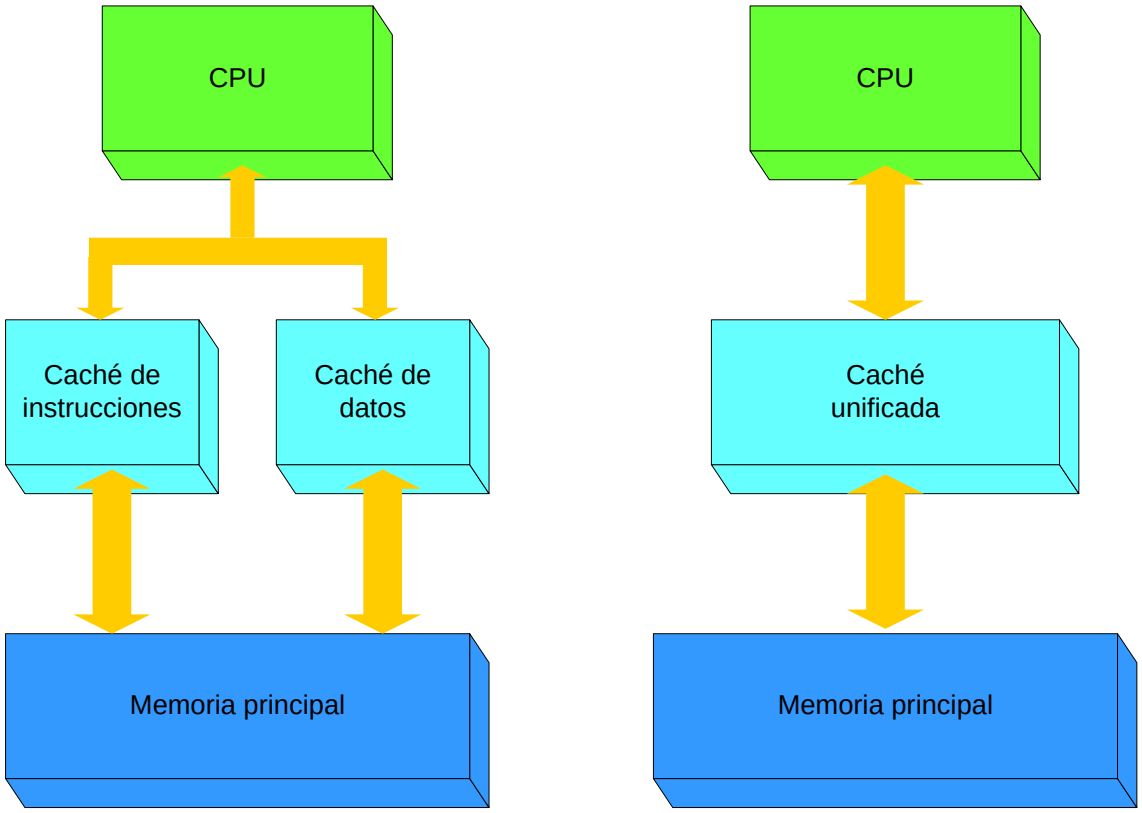
- Se añade un segundo nivel de caché fuera del chip (L2) con un tiempo de acceso menor que el de la memoria principal.
- Cuando ocurre un fallo en la caché primaria (L1), se accede a la caché secundaria (L2) para buscar los datos.

  - Si están allí se reduce la penalización de fallo.

  - Si no están accedemos a la memoria principal.
- Normalmente la L1 es de pequeño tamaño y con asociatividad 2 o 4, mientras que la L2 es grande y con asociatividad 2, como mucho.



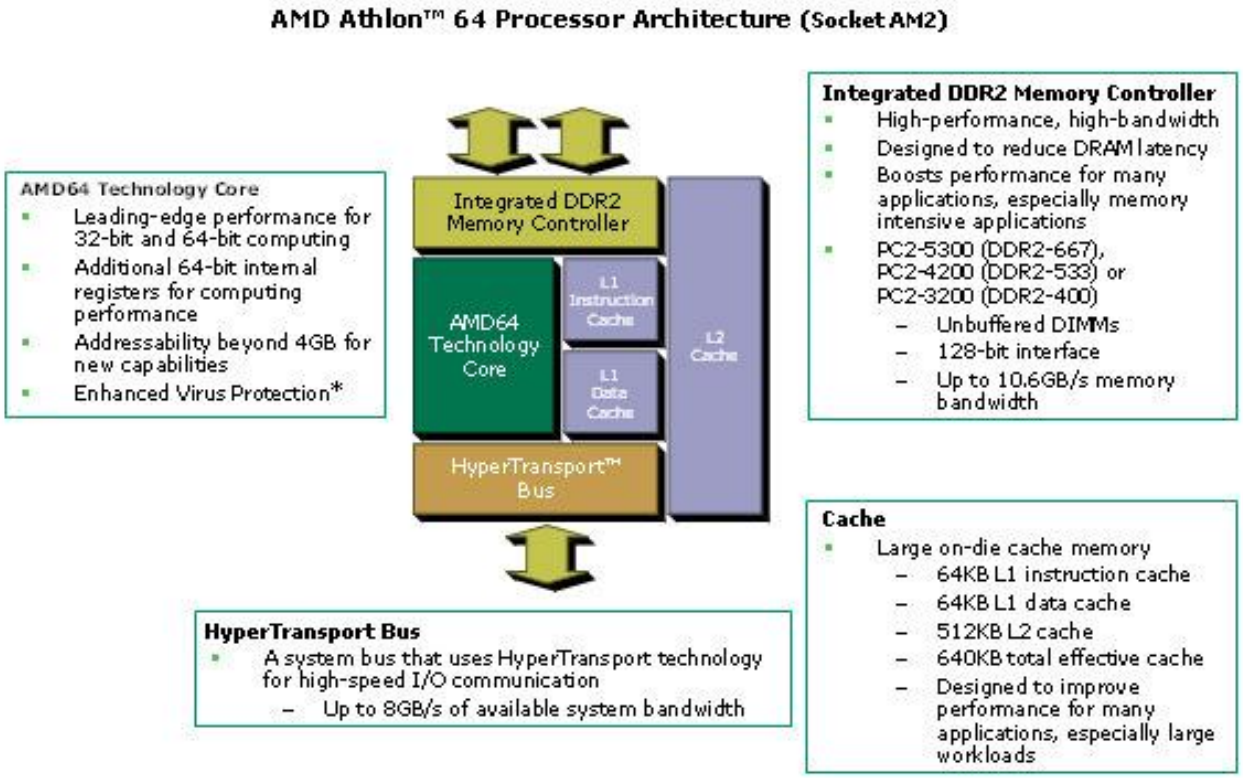
Memorias cache multinivel (II)



Índice

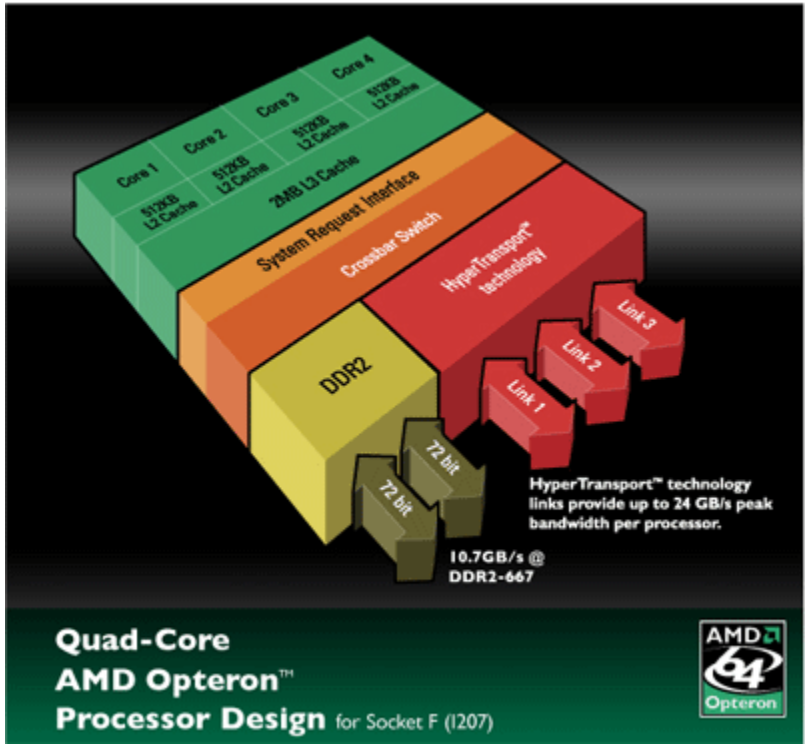
Memorias caché en microprocesadores actuales (I)

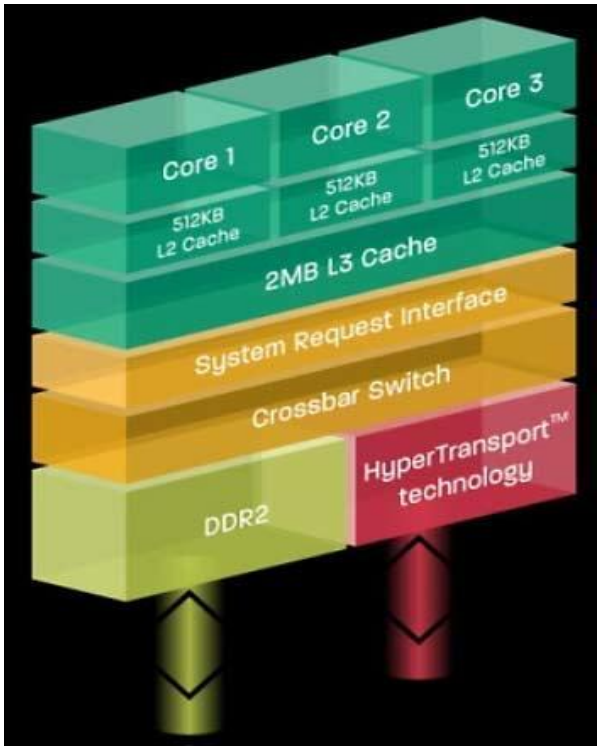
AMD Athlon



Memorias caché en microprocesadores actuales (II)

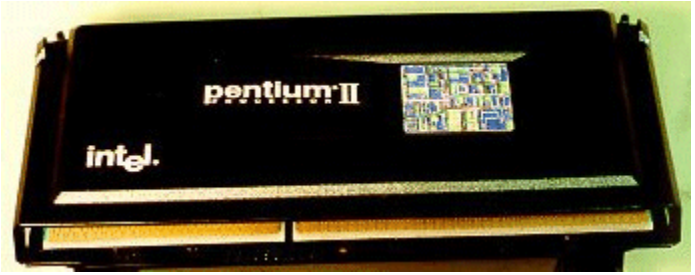
AMD Opteron - Phenom X3





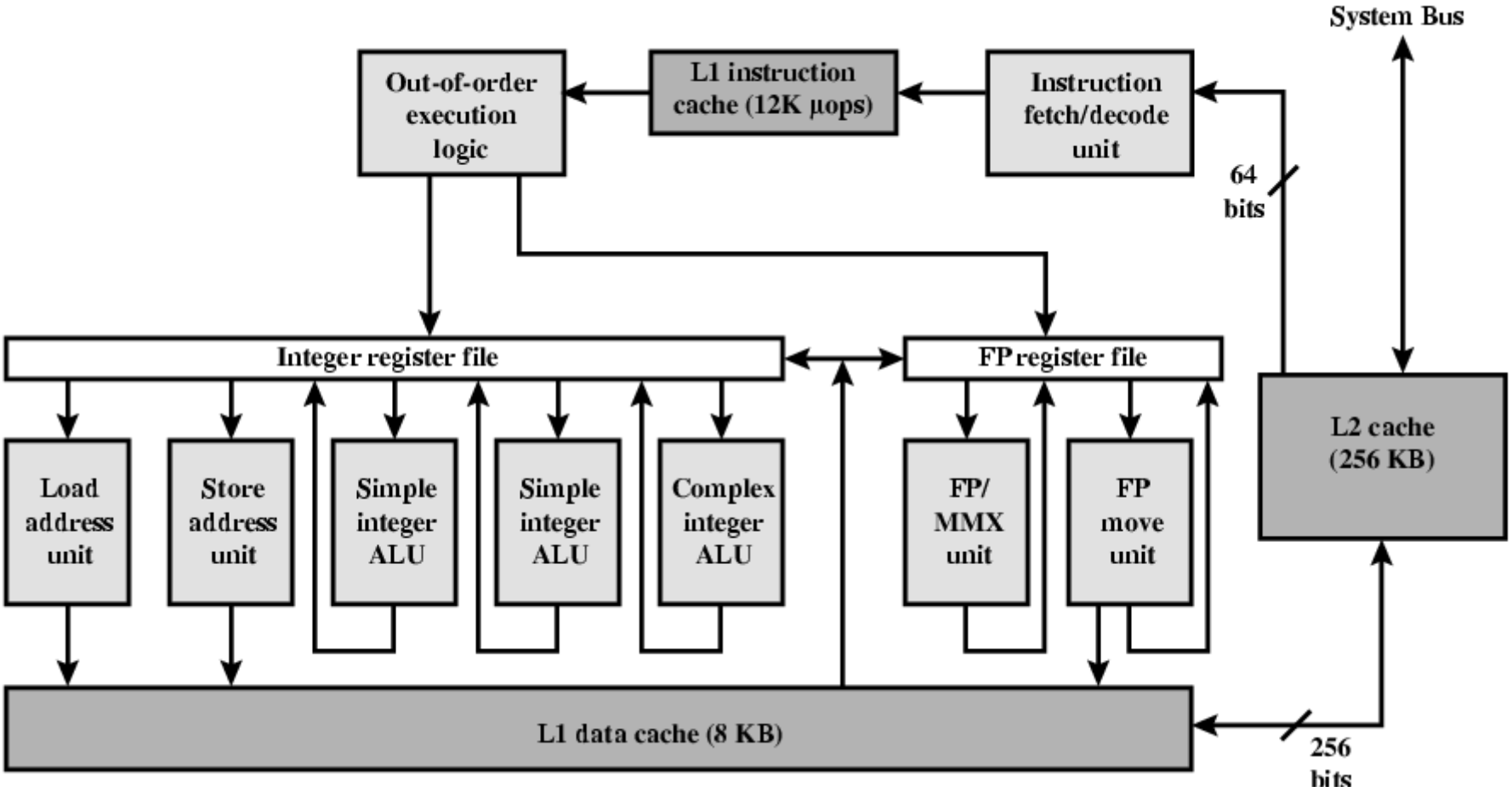
Memorias caché en microprocesadores actuales (III)

Intel Pentium II



Memorias caché en microprocesadores actuales (IV)

Pentium 4



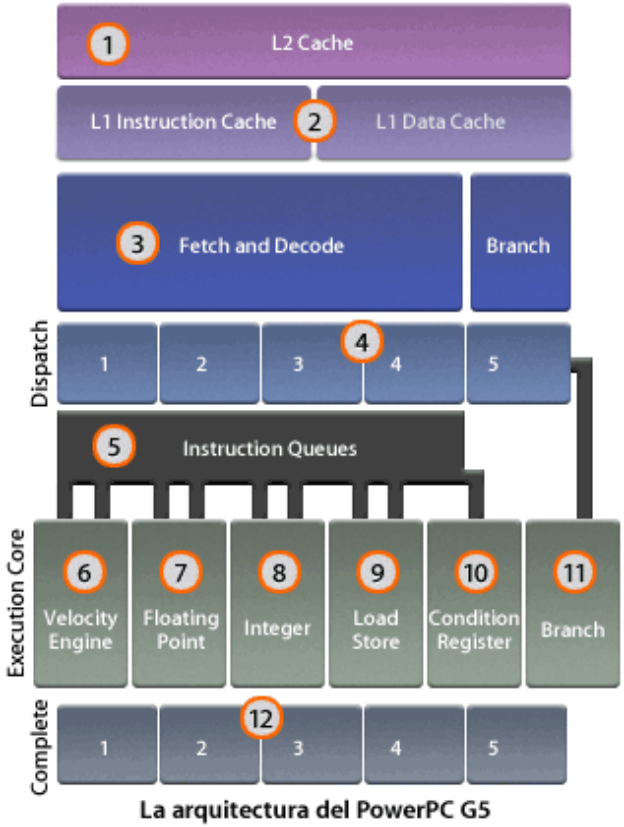
Memorias caché en microprocesadores actuales (V)

Intel Itanium 2

- L1I +L1D
  - 16k bytes por caché
  - Asociativa por conjuntos: 4 líneas
- L2
  - 256k bytes
  - Asociativa por conjuntos: 8 líneas
- L3
  - 6 M bytes
  - Asociativa por conjuntos: 12 líneas

Memorias caché en microprocesadores actuales (VI)

PowerPC G5



La arquitectura del PowerPC G5

- Caché L2

512 K de caché L2 facilitan al núcleo de ejecución acceso ultrarrápido a 64 MBps a datos e instrucciones.

- Caché L1

Dotada de mapeado directo de 64 K a 64 GBps.

UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA

Microprocesadores para comunicaciones

Escuela Técnica Superior de Ingenieros de Telecomunicación

Organización y estructura de la memoria cache