Universidad de Jaén

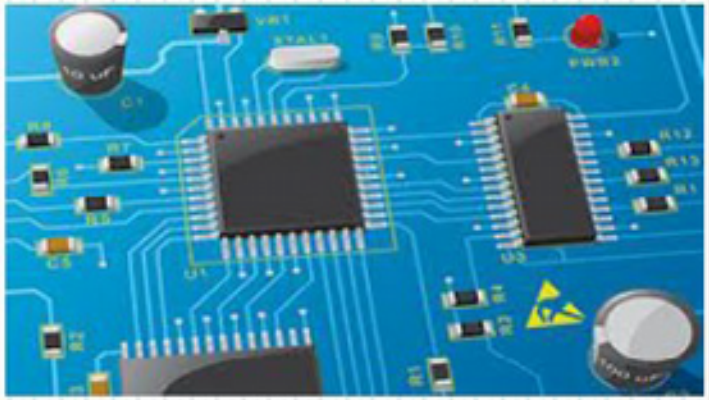
Compromiso con la sociedad

25 AÑOS DE UNIVERSALIDAD 1993-2018

T3. Sistema de Memoria:

T3.3 Introducción a la Memoria Caché

FUNDAMENTOS DE ARQUITECTURA DE COMPUTADORES



Departamento de Informática

Contenido del capítulo

- Introducción y Tecnología de la memoria caché
- Principios de funcionamiento
- Conceptos generales y posicionamiento
- Diseño de la memoria caché:
  - Diseño de la estructura de las entradas de caché
  - Política de ubicación
  - Direccionamiento directo
- Bibliografía

Contenido del capítulo

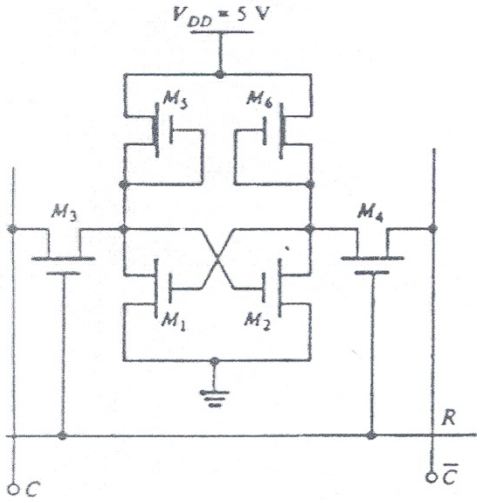
- Introducción y Tecnología de la memoria caché
- Principios de funcionamiento
- Conceptos generales y posicionamiento
- Diseño de la memoria caché:
  - Diseño de la estructura de las entradas de caché
  - Política de ubicación
  - Direccionamiento directo
- Bibliografía

Introducción

- Es una pequeña memoria construida con chips de mem. estática (SRAM) de alta velocidad y costosos.
- Situado entre microprocesador y memoria DRAM.
- Objetivo: aumentar el rendimiento en los accesos que el microprocesador debe realizar a la MP.
- Se pretende que el microprocesador pueda operar la mayor parte del tiempo desde la caché
- La caché contendrá palabras de código y datos de la memoria principal que el microprocesador utiliza frecuentemente.

Tecnología de la Memoria Caché

- Se fundamentan en un biestable como celda memoria que se construye a partir de 4 o 6 transistores.
- Más espacio y mayor coste de integración que DRAM.
- Consumo más elevado que DRAM, aunque más rápidas.
- No necesitan ser refrescadas.
- No necesitan tiempo de precarga.



Contenido del capítulo

- Introducción y Tecnología de la memoria caché
- Principios de funcionamiento
- Conceptos generales y posicionamiento
- Diseño de la memoria caché:
  - Diseño de la estructura de las entradas de caché
  - Política de ubicación
  - Direccionamiento directo
- Bibliografía

Funcionamiento: localidad

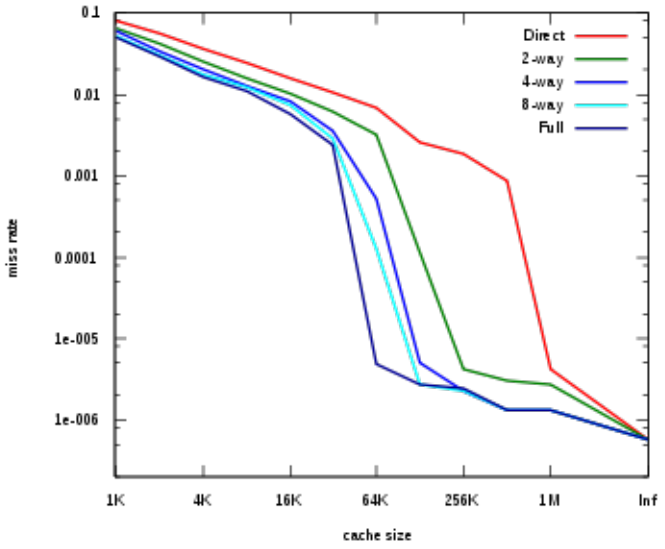
- Éxito funcionamiento de caché: principio de localidad:
  - Localidad **temporal**:
    - Si se accede a una dirección determinada es muy posible que vuelva a ser referenciada en poco tiempo.
    - Instrucciones: **Bucles**, subrutinas, etc.
    - Datos: Uso de subconjunto de variables
  - Localidad **espacial**:
    - Si se accede a una dirección determinada es muy posible que se acceda posteriormente a las direcciones cercanas a esa.
    - El 90% del tiempo de ejecución se consume en aprox. el 10% del código.
    - Instrucciones: Ejecución secuencial.
    - Datos: **Vectores**, etc.

Funcionamiento: Frec. De Aciertos

- Cada acceso del microprocesador a memoria es interpretado por el controlador de la caché:
  - Si el dato está disponible se envía al procesador: se dice que se ha producido un **acierto**,
  - Si no lo está, se traspasa la petición a la memoria principal: se ha producido un **fallo**.
- En el último caso, la caché almacena el dato pues será muy probable que vuelva a ser referenciado (Política de extracción).
- Objetivo de la caché: **anticiparse** a las peticiones del microprocesador de manera que se reduzcan los fallos.

Funcionamiento: ¿El tamaño importa?

- ¿Mayor capacidad implica mayor probabilidad de aciertos?
  - Coste:
    - Alto precio SRAM.
    - Más espacio → mayor complejidad.
  - Eficiencia → Principio de localidad:
    - Almacenar en la caché código y datos que no están cerca a los referenciados recientemente no mejora rendimiento:
      - Al salirse fuera del entorno de vecindad, no es probable que sean requeridos.
      - Estamos desaprovechando memoria de alto coste inútilmente.



Contenido del capítulo

- Introducción y Tecnología de la memoria caché
- Principios de funcionamiento
- Conceptos generales y posicionamiento
- Diseño de la memoria caché:
  - Diseño de la estructura de las entradas de caché
  - Política de ubicación
  - Direccionamiento directo
- Bibliografía

Conceptos generales

- **Terminología:**
  - *Palabra:* unidad máxima de información de en un registro de la CPU.
  - *Unidad direccionable*: porción mínima de datos de memoria principal a la que se puede acceder. Su tamaño suele coincidir con la *palabra*.
  - *Direcciones de memoria*: es una cadena de bits con la que nos referimos a cualquiera de las posiciones existentes en la memoria principal. Su longitud depende por tanto del número de *unidades direccionables*.
  - *Bloque de memoria*: las *unidades direccionables* se agrupan en bloques de mayor tamaño para su gestión con la memoria caché.
  - *Línea de caché*: es idéntico a un bloque de memoria principal, pero en este caso nos referimos a datos que están en la caché.
  - *Marco de bloque*: se utiliza como un sinónimo de la *línea de caché*.
  - *Conjunto*: agrupación de *líneas de caché* para su uso en memorias caché de tipo asociativo (que se explicarán más adelante).

Conceptos generales (2)

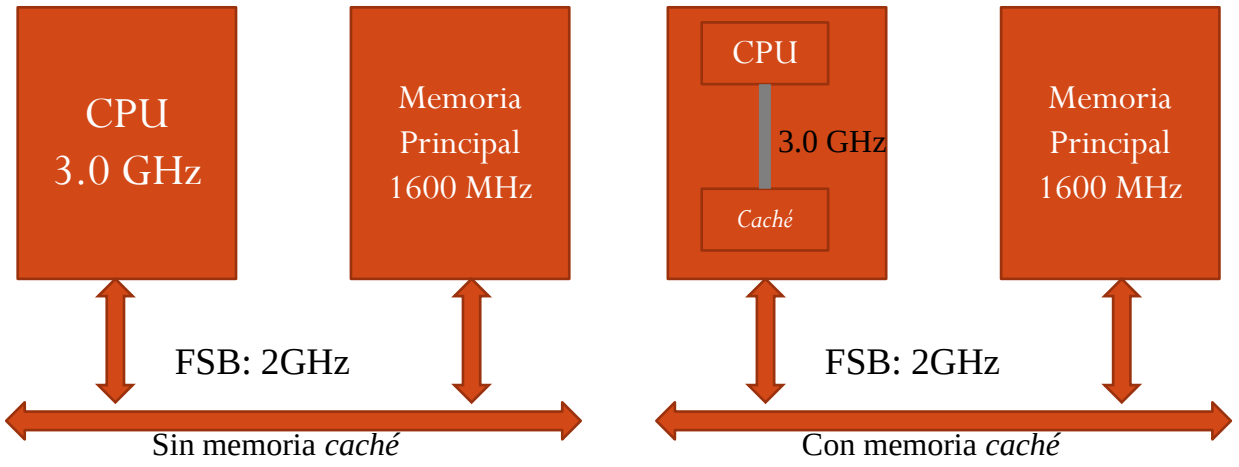
- Función de correspondencia:
  - Determina las posibles líneas de la caché en las que se puede ubicar un bloque dado de Memoria Principal
  - En dicho bloque se encuentra la palabra que ha sido referenciada por el programa. Por tanto ha de llevarse a memoria caché (principio localidad).
- Algoritmo de sustitución:
  - Es necesario hacer espacio para ubicar un nuevo bloque.
  - Determina la línea de caché que hay que liberar cuando está llena

Conceptos generales (3)

- Política de escritura:
  - Forma de mantener la coherencia entre memoria caché y memoria principal
  - Entra en juego cuando se realizan modificaciones (escrituras) en la caché y de actualizarse la MP
- Política de búsqueda de bloques:
  - Determina la causa que desencadena traer un bloque a la caché (ej. un fallo en la referencia)
  - Se refiere principalmente al concepto de "anticipación"
- Cachés independientes para datos e instrucciones:
  - Frente a cachés unificadas, éstas pueden ser más útiles y mejorar el rendimiento global del sistema.

Posicionamiento de la Memoria Caché

- Actualmente la memoria caché (L1 y L2) está integrada con el microprocesador funcionando a su misma velocidad y teniendo una conexión mas directa y ancha (bus) que antes.
- Esto permite ser mucho más eficiente todavía que la memoria principal DRAM:
  - Más cerca del procesador
  - Evita competir por recursos del bus
  - No hay latencia por precarga, etc.

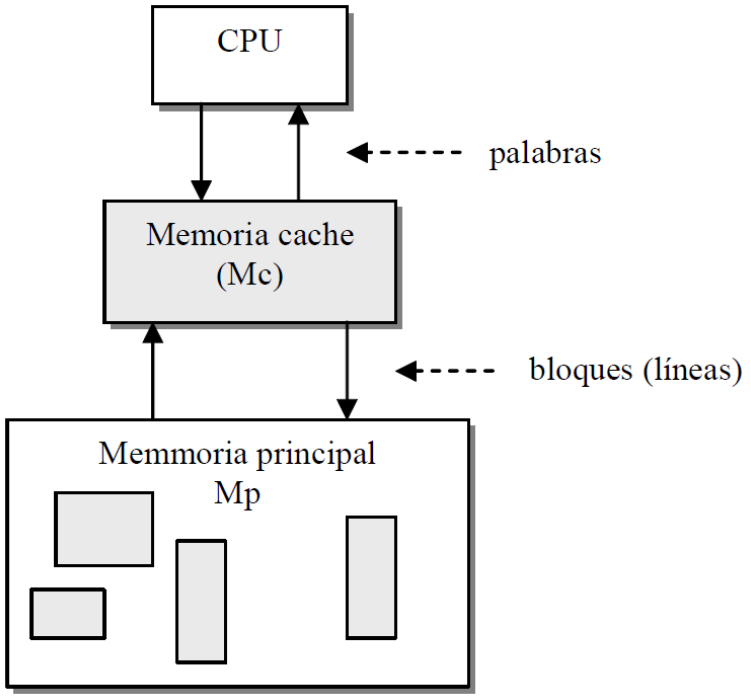


Contenido del capítulo

- Introducción y Tecnología de la memoria caché
- Principios de funcionamiento
- Conceptos generales y posicionamiento
- Diseño de la memoria caché:
  - Diseño de la estructura de las entradas de caché
  - Política de ubicación
  - Direccionamiento directo
- Bibliografía

Diseño y Estruct.: Longitud de datos

- Para implementar el mecanismo de actualización de la caché se divide la memoria principal en **bloques** de un *múltiplo de palabras*
- La caché se compone de marcos de bloque o **líneas** de *igual tamaño*.
- (*) El bloque será la unidad de intercambio de información entre la memoria principal y la caché.
- (*) Entre la caché y la CPU sigue siendo la palabra (registros).



Diseño y Estruct. entradas de caché

- Cada entrada (posición) de caché, contendrá la siguiente información:
  - Bit de validez: si la información pertenece al programa en ejecución.
  - Bit de "suciedad" (dirty bit): si la información se ha modificado en algún momento.
  - Etiqueta: índice para encontrar los datos.
  - Bloque de datos (*línea* de caché): datos tomados de M.P.

Diseño y Estruct. entradas de caché

- La memoria caché está formada por la agrupación de varias líneas, todas ellas con la misma estructura.
- El tamaño del bloque de datos coincide con el de la M.P.

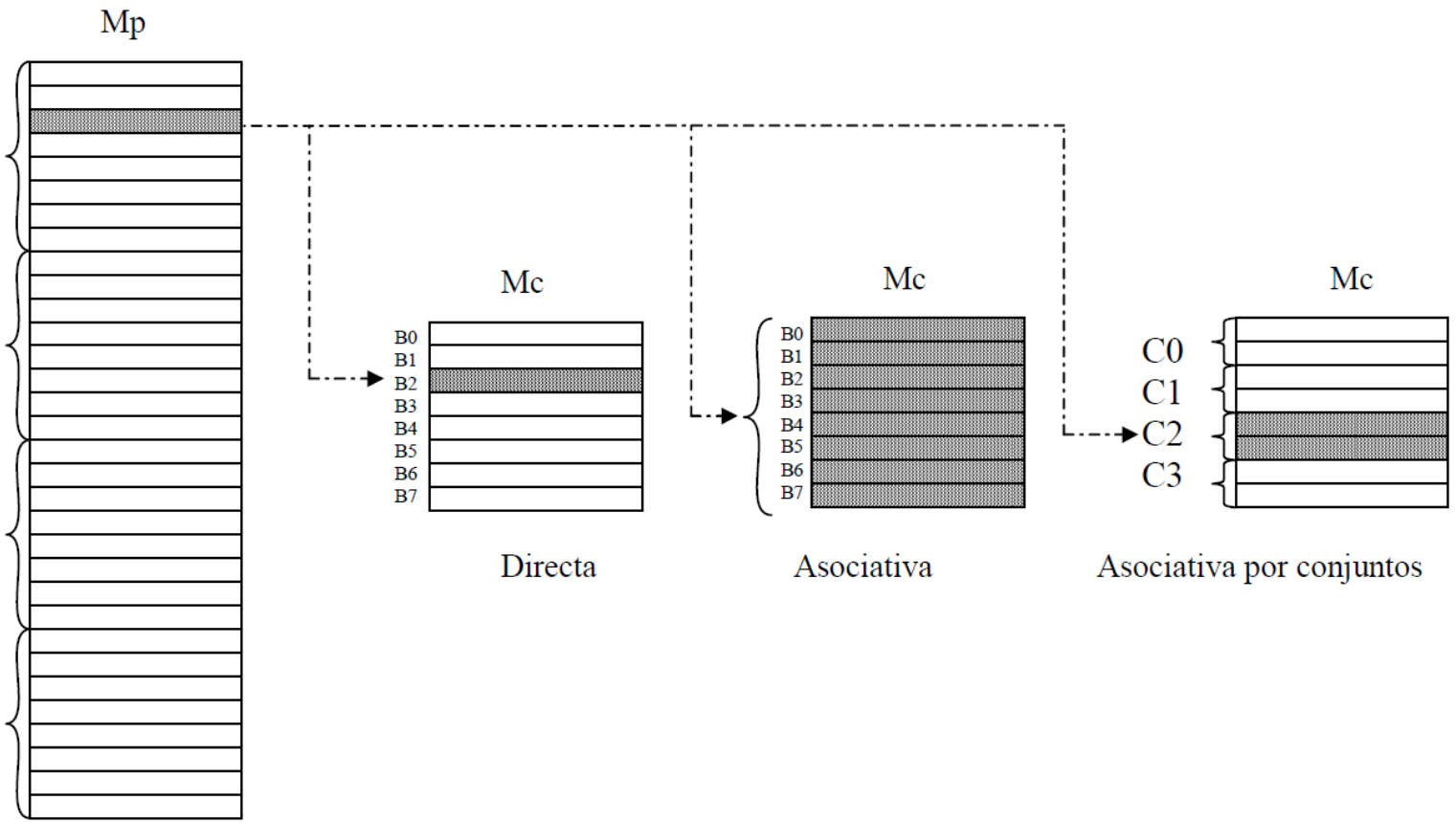
Contenido del capítulo

- Introducción y Tecnología de la memoria caché
- Principios de funcionamiento
- Conceptos generales y posicionamiento
- Diseño de la memoria caché:
  - Diseño de la estructura de las entradas de caché
  - Política de ubicación
  - Direccionamiento directo
- Bibliografía

Política de Ubicación

- Existen tres funciones de correspondencia para definir la posible ubicación de un bloque de MP en la memoria caché:
- **Directa:** un bloque de MP sólo puede ubicarse en una línea de la caché.
- **Asociativa:** un bloque puede ubicarse en cualquier línea
- **Asociativa por conjuntos:** es un compromiso entre las dos anteriores

Política de Ubicación (2)

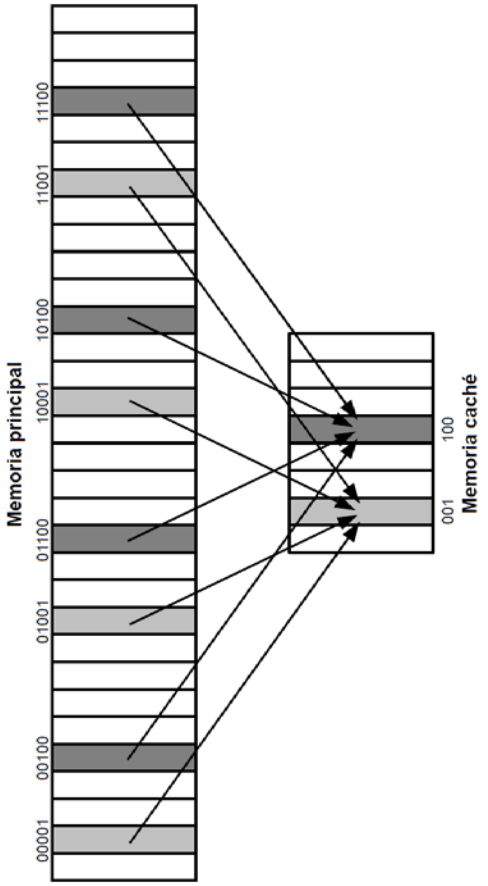


Contenido del capítulo

- Introducción y Tecnología de la memoria caché
- Principios de funcionamiento
- Conceptos generales y posicionamiento
- Diseño de la memoria caché:
  - Diseño de la estructura de las entradas de caché
  - Política de ubicación
  - Direccionamiento directo
- Bibliografía

Ubicación: Correspondencia directa

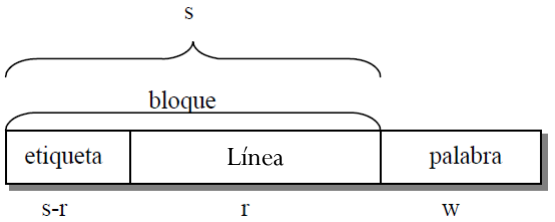
- Dado un bloque de MP, **SOLAMENTE** puede estar en una línea fija de caché:
  - Con **N** líneas de caché, el bloque **X** de memoria principal le corresponde la línea **Y** de caché.
  - **Y = X módulo N**,
- Existen varios bloques de MP para cada línea de caché
- El bloque de MP que en un momento dado estará guardado en la caché vendrá determinado por los bits de **"etiqueta"**.



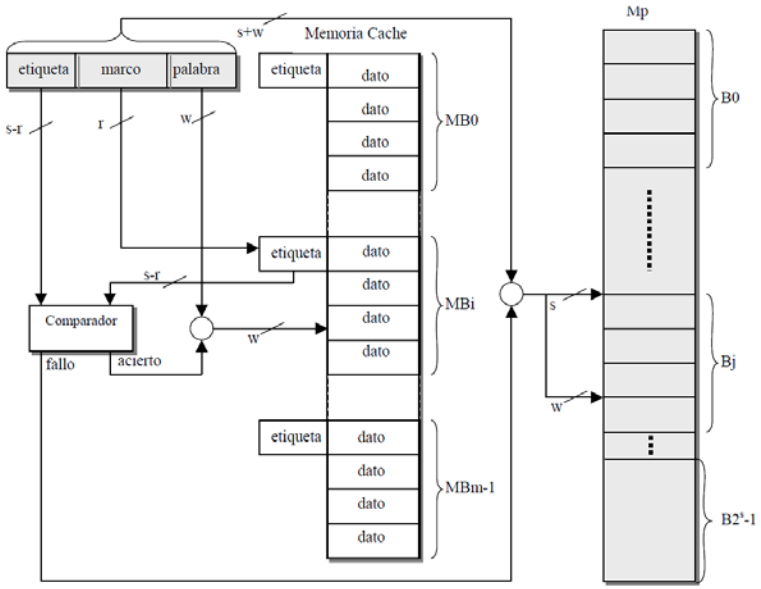
Ubicación: Correspondencia directa

- 2^x palabras en MP
  - X = #bits dirección de memoria
- 2^s bloques de MP
- 2^w palabras/bloque
- 2^r líneas en MC
  - 2^r = N, N = #líneas
- 2^{s-r} veces está contenida la MC en MP

Dirección de Memoria



Diseño de la M.C.



Ejemplo de Direccionamiento Directo

- Imaginemos un sistema con una memoria principal de 1MiB, junto con una memoria caché de 8 KiB. La longitud de palabra es de 32 bits y los bloques se agrupan de 16 en 16 palabras.

- Calculamos el número de palabras (unidades direccionables) de Memoria Principal (MP):
  - MP = 1MiB; 1 MiB = $1 \cdot 2^{20} \cdot 2^3$ bits = 2^{23} bits
  - Si hay 32 bits por palabra → $2^{23}/2^5 = 2^{18}$ posiciones o palabras totales en la Memoria Principal.
- Debemos conocer el número de bloques en función del número de palabras que almacene:
  - En el ejemplo hay 16 palabras por bloque
  - 2^{18} palabras / 2^4 palabras en un bloque = 2^{14} bloques en M.P.

Ejemplo de Direccionam. Directo (2)

- El número de líneas de caché es fijo y depende de la capacidad:
  - MC = 8KiB = $2^3 \cdot 2^{10} \cdot 2^3$ bits = 2^{16} bits
  - 2^{16} bits totales en MC / 2^5 bits por palabra = 2^{11} palabras
  - 2^{11} palabras en MC / 2^4 palabras por bloque (o línea) = 2^7 líneas de caché, cada una con 16 palabras (como la MP).
- Como hay más bloques de MP que de caché se pueden asignar un total de:
  - 2^{14} bloques MP / 2^7 líneas caché = 2^7 bloques de MP a cada línea.
- En este caso la etiqueta indica el identificador del bloque de MP:
  - Cada bloque de MP solo puedo ir a una línea,
  - **¡pero una línea puede recibir varios bloques!**
  - En concreto, cada línea recibe hasta 2^7 bloques por lo que son necesarios 7 bits para definir cada uno de ellos.

Ejemplo de Direccionam. Directo (3)

- En este ejemplo, las direcciones deben ser siempre de 18 bits, ya que hay 2^{18} palabras en la memoria principal
- Dirección de petición de datos:
  - 0xD1FA; Se traduce a →

Marca	Bloque	Palabra
1101000	1011111	1010

- Para la dirección dada, se realizan los siguientes pasos
  - Comprobar primero el índice (bloque o línea de caché): 0x5F
  - Validez = 1: compara etiqueta 0x68 indica bloque de MP
    - Si éstos coinciden, se lee la palabra que se encuentra en la posición 0x0A
    - En otro caso, si el bit de suciedad es 1 se actualiza la MP y se trae el bloque completo desde MP, actualizando flags: validez = 1; suciedad = 0.
  - Validez = 0, se trae el bloque completo desde MP. Se actualizan también los flags: Validez = 1; Suciedad = 0.

Contenido del capítulo

- Introducción y Tecnología de la memoria caché
- Principios de funcionamiento
- Conceptos generales y posicionamiento
- Diseño de la memoria caché:
  - Diseño de la estructura de las entradas de caché
  - Política de ubicación
  - Direccionamiento directo
- Bibliografía

Bibliografía

- Patterson y Hennessy: Estructura y Diseño de Computadores. Capítulo 5.

- Murdocca y Heuring: Principios de Arquitectura de Computadoras: Capítulo 7.

- Memoria del computador. Disponible en http://www.slideshare.net/Sofylutqm/memoria-del-computador