Introducción a la Computación

Capitulo 4

Memoria Cache

Características generales

- Ubicación
- Capacidad
- Unidad de transferencia
- Método de Acceso
- Prestaciones
- Dispositivo Físico
- Características Físicas
- Organización

Ubicación

- CPU
- Interna
- Externa

Capacidad

- En bytes o palabras
  - La unidad natural de organización
- Tamaño de la palabra
  - 8, 16, 32 ... bits

Unidad de transferencia

- Interna
  - Definida por el ancho del bus de datos
- Externa
  - Bloque mucho mas grande que una palabra
- Unidad Direccionable
  - Locación mas pequeña que puede ser direccionada
  - Internamente es una palabra pero puede ser bytes
  - Clusters en HD

Métodos de Acceso (1)

- Secuencial
  - Unidades de datos denominados “Registros”
  - Comienza desde el principio y avanza secuencialmente
  - Tiempo de acceso depende de la ubicación del dato buscado y de la posición actual
  - Ej. Cinta
- Directo
  - Bloques individuales que tienen dirección única
  - Se accede a la bloque y dentro de él se busca secuencialmente
  - Tiempo de acceso es nuevamente variable
  - Ej. Disco

Métodos de Acceso (2)

- Aleatorio (Random)
  - Cada posición de memoria tiene una única dirección
  - El tiempo es el mismo para cada posición
  - Ej. RAM
- Asociativa
  - El dato es recuperado mediante una comparación de los contenidos comunes de todas las celdas a la vez
  - El tiempo de acceso es fijo
  - Ej. Cache

Prestaciones

- Tiempo de acceso
  - Tiempo entre que se presenta la dirección a la memoria y se memoriza o se hace disponible
- Tiempo de Ciclo de memoria
  - La memoria necesita de un tiempo para recuperarse entre accesos
  - Tiempo de acceso + recuperación
- Velocidad de transferencia
  - Velocidad a la cual se pueden transferir los datos

Soportes Físicos

- Semiconductor
  - RAM
- Magnético
  - Discos y Cinta
- Óptico
  - CD y DVD

Características Físicas

- Perdurabilidad
  - Volatilidad
  - Borrable o no
- Consumo de energía

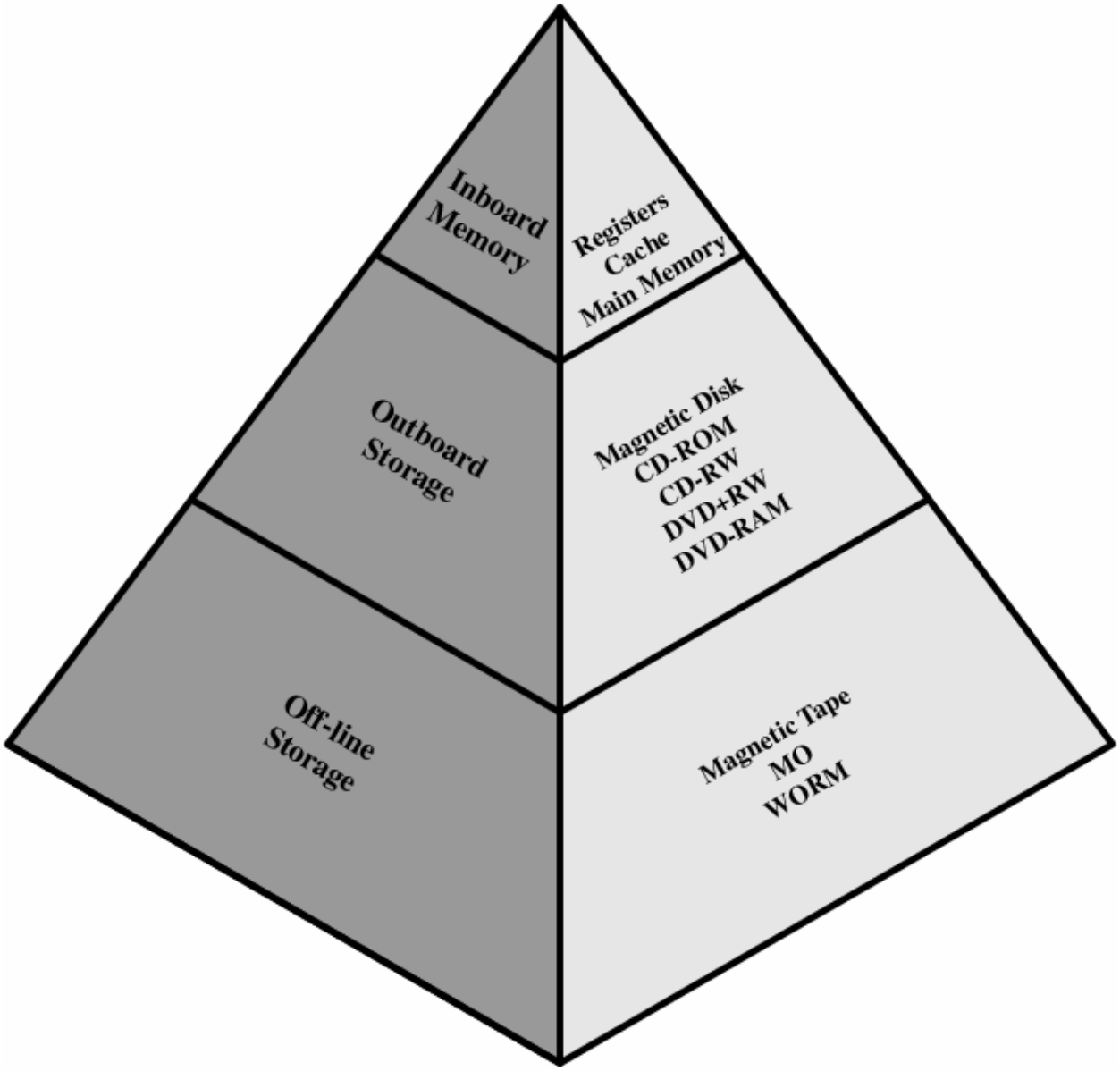
Organización

- Disposición de los bits en las palabras
- No siempre es obvia
- Ej. Interlineado

Jerarquía de la memoria

- Registros
  - En CPU
- Memoria Interna o Principal
  - Puede incluir uno a o mas niveles de cache
  - “RAM”
- Memoria Externa
  - Almacenamiento permanente

Jerarquía de la memoria - Diagrama



Costo, Capacidad y Velocidad

- A menor tiempo de acceso
  - —Aumenta el costo por bit
- A mayor capacidad
  - —Menor costo por bit
- A mayor capacidad
  - —Mayor tiempo de acceso

Lista Jerárquica

- Registros
- Cache L1
- Cache L2 ...
- Memoria principal
- Cache de Disco
- Disco
- Unidades Ópticas
- Cinta

¿Si se quiere velocidad?

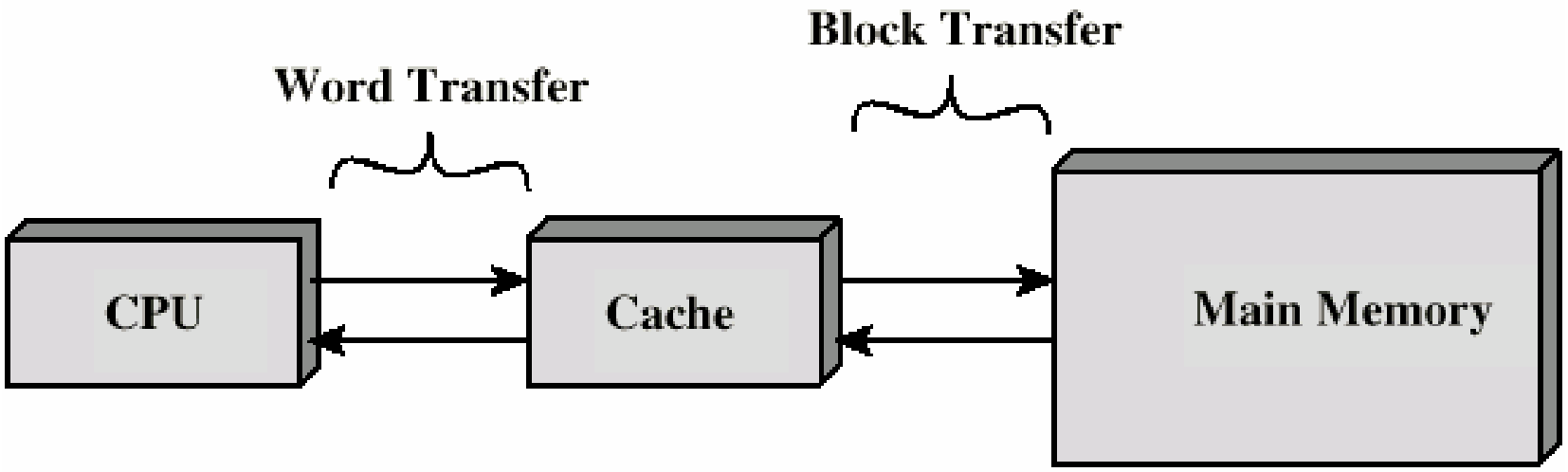
- Es posible construir una computadora que use solo RAM estática
- Esta sería muy rápida
- No necesitaría cache
- Costaría demasiado

Localidad de las referencias

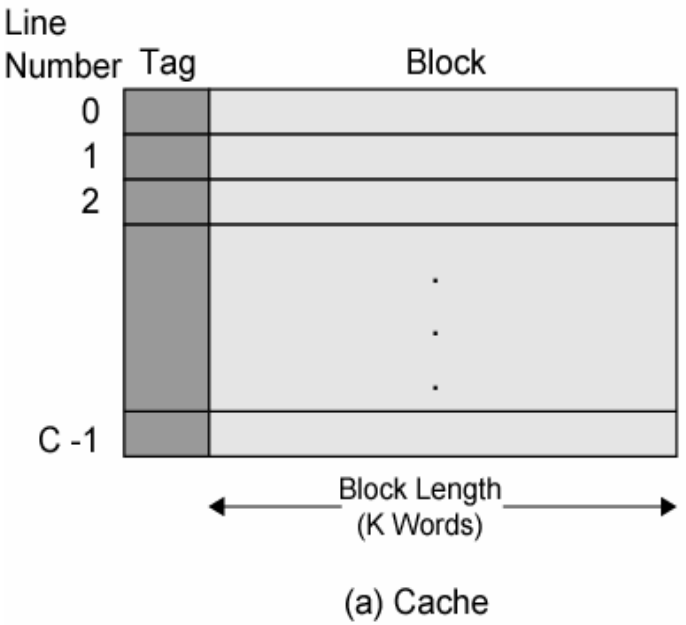
- Durante la ejecución de un programa, las direcciones de las instrucciones tienden a estar agrupadas
  - Ej. Bucles
- Lo mismo ocurre con los datos
  - Ej. Matrices

Cache

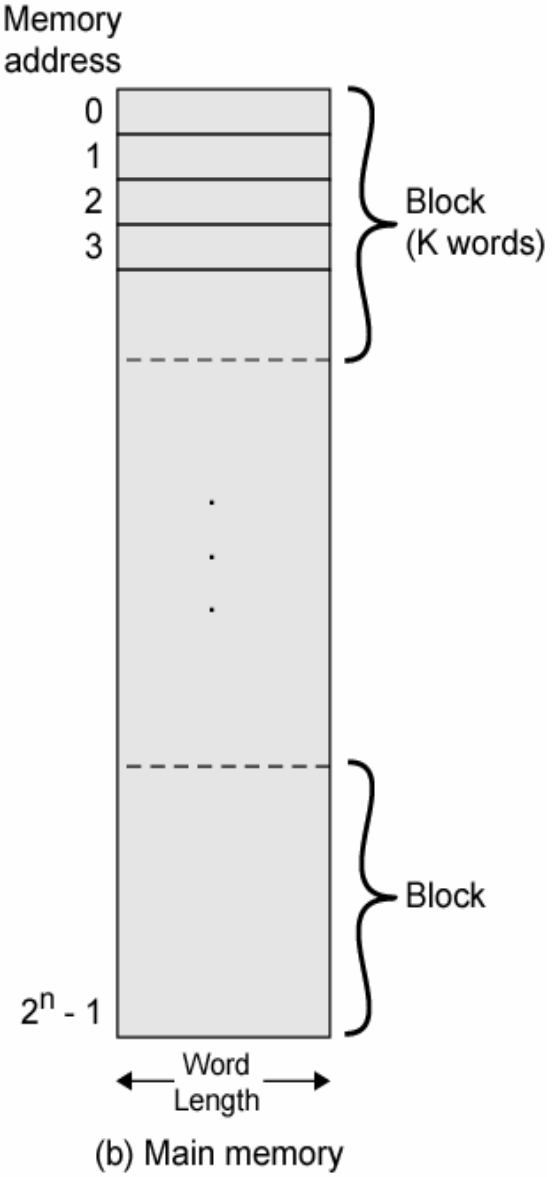
- Pequeña porción de memoria rápida
- Entre la memoria principal y la CPU
- Puede estar en CPU o como módulo aparte



Estructura del Cache/Memoria Principal



(a) Cache

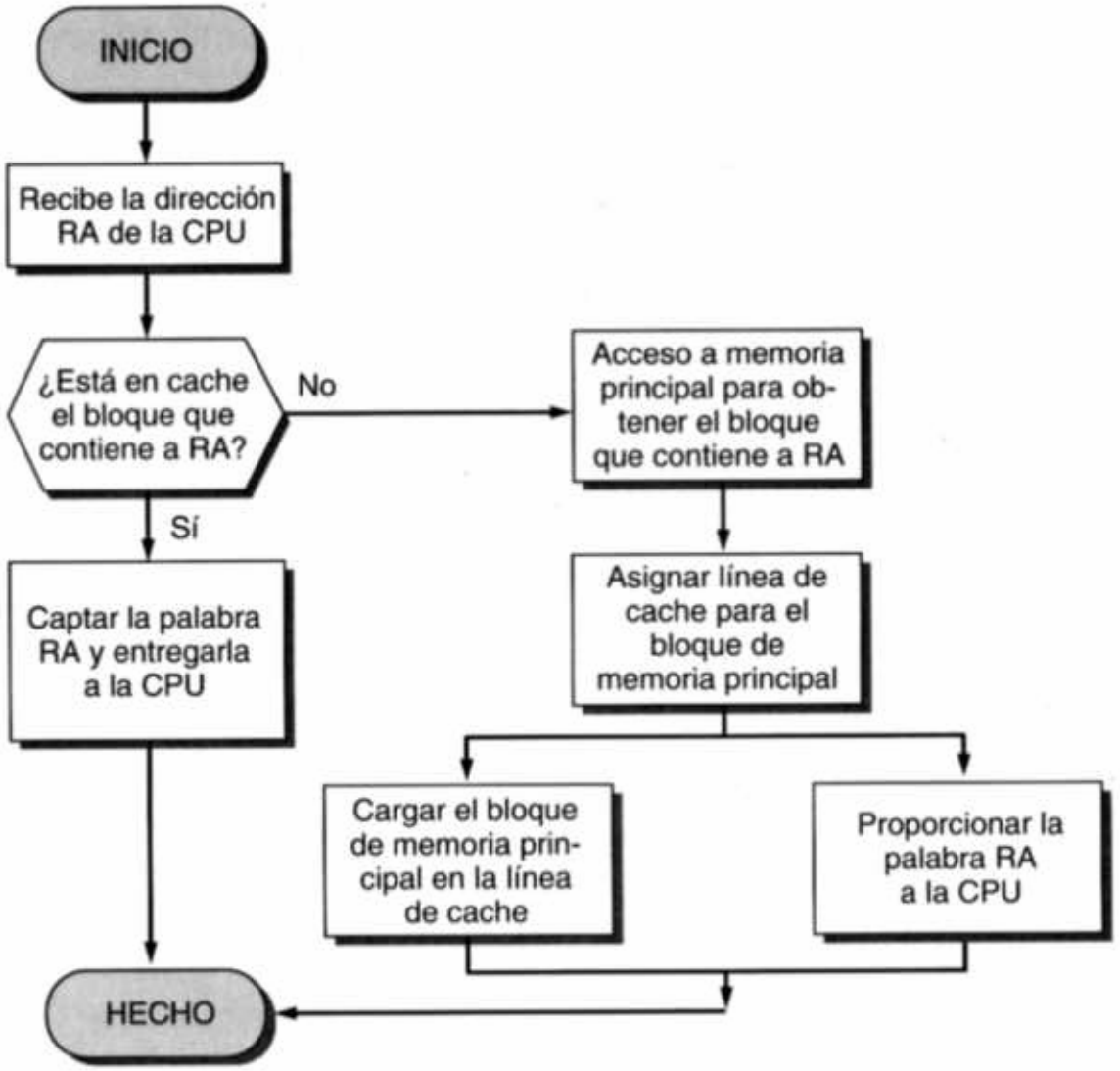


(b) Main memory

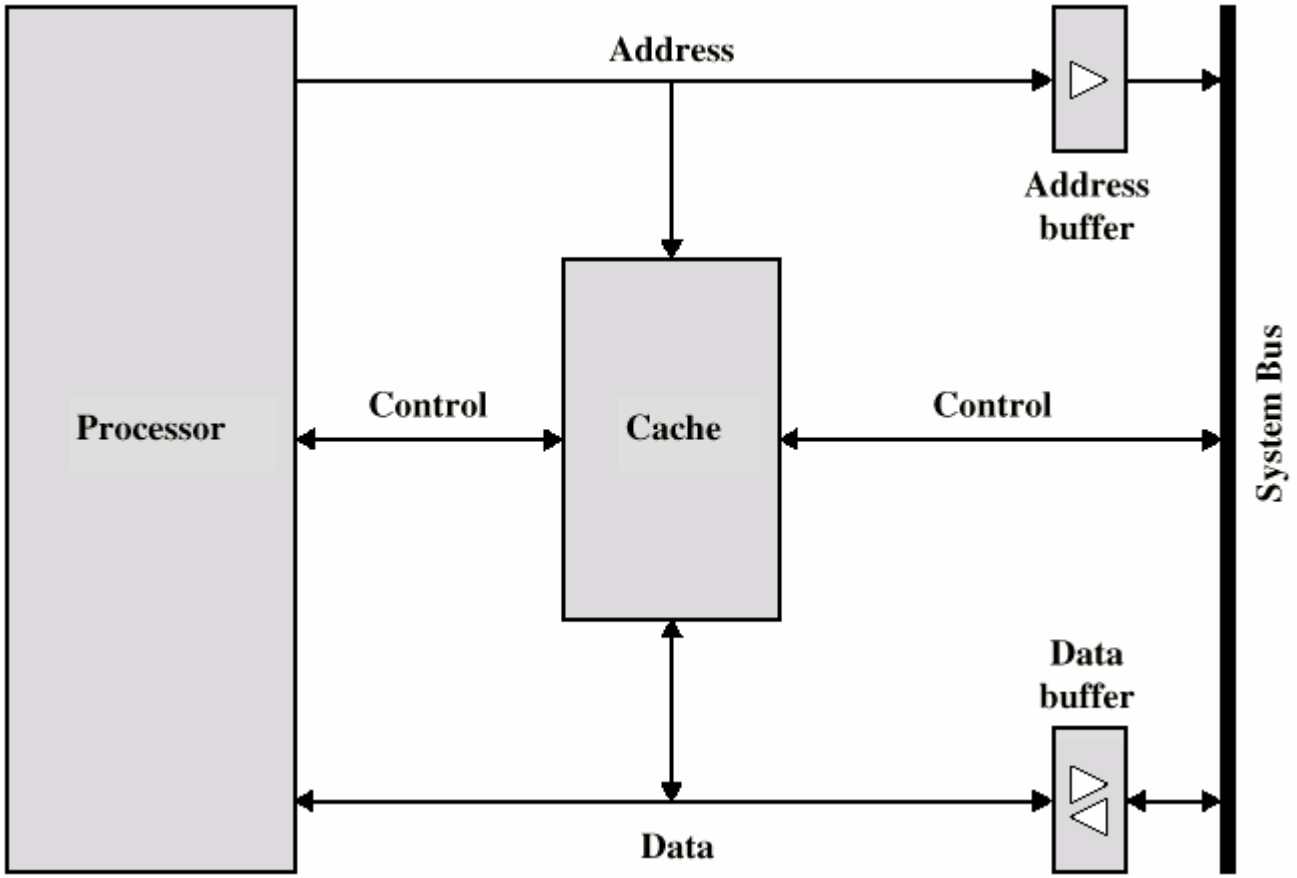
Operación del Cache - Principios

- CPU pide el contenido de una posición de memoria
- Verifica el cache para ver si se encuentra ahí
- Si es así, lo lee de cache (rápido)
- Si no, se transfiere el bloque de memoria donde se encuentra la posición requerida, al cache
- Se entrega la posición a la CPU
- Cache incluye etiquetas que identifican cual bloque de memoria principal esta en cada línea de cache

Operación - Diagrama de Flujo



Organización Típica del Cache



Diseño del Cache

- Tamaño
- Función de correspondencia
- Algoritmo de sustitución
- Política de escritura
- Tamaño de bloque
- Número de caches

Tamaño

- Costo
  - Mucho cache es costoso
- Velocidad
  - Mas cache hace mas rápido el funcionamiento (hasta cierto punto)
  - Verificar el cache toma tiempo, cuando mas grande es, mas tiempo toma.

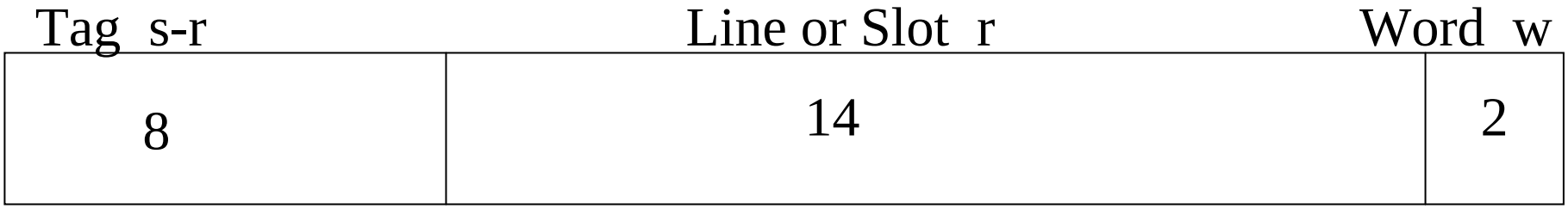
Función Correspondencia (Mapeo)

- El ejemplo se basa en:
  - Cache de 64kB
  - Bloque de Cache de 4 bytes
    - 16k líneas (2^{14}) de 4 bytes
  - 16MB de memoria principal
  - Direcciones de 24 bit
    - (2^{24}=16M)

Mapeo directo

- Cada bloque de memoria principal se corresponde con una sola línea de cache
  - Ej. Si un bloque esta en cache, este debe estar en un lugar específico
- Direcciones de dos partes
- Menos significativa **w** bits identifica la palabra en el bloque
- Mas significativa s bits especifica el bloque de memoria
- La MSB esta dividida en un campo línea de cache r y una etiqueta s-r (mas significativa)

Estructura del direccionamiento del Mapeo Directo



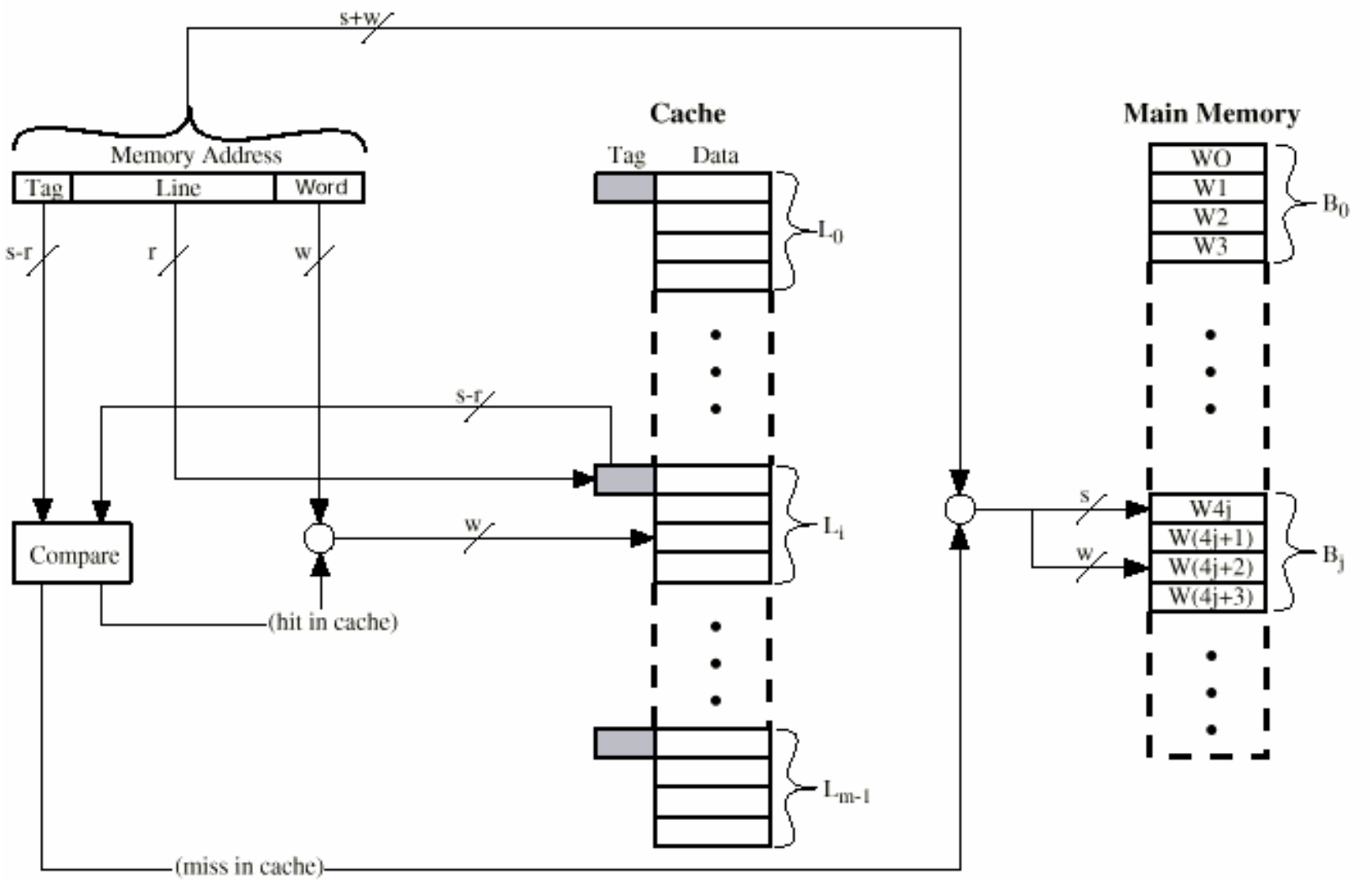
- Direcciones de 24 bit
- Identificador de palabra de 2 bit (bloque de 4 bytes)(w)
- Identificador de bloque de 22 bit(s)
  - Línea de 14 bit (r) => 2^r = m líneas de cache
  - Etiqueta de 8 bit (=22-14) (s-r)
- No hay 2 bloques que se asignen en la misma línea que tengan el mismo numero de etiqueta
- Busca la línea y verifica la etiqueta

Tabla de líneas de cache del Mapeo Directo

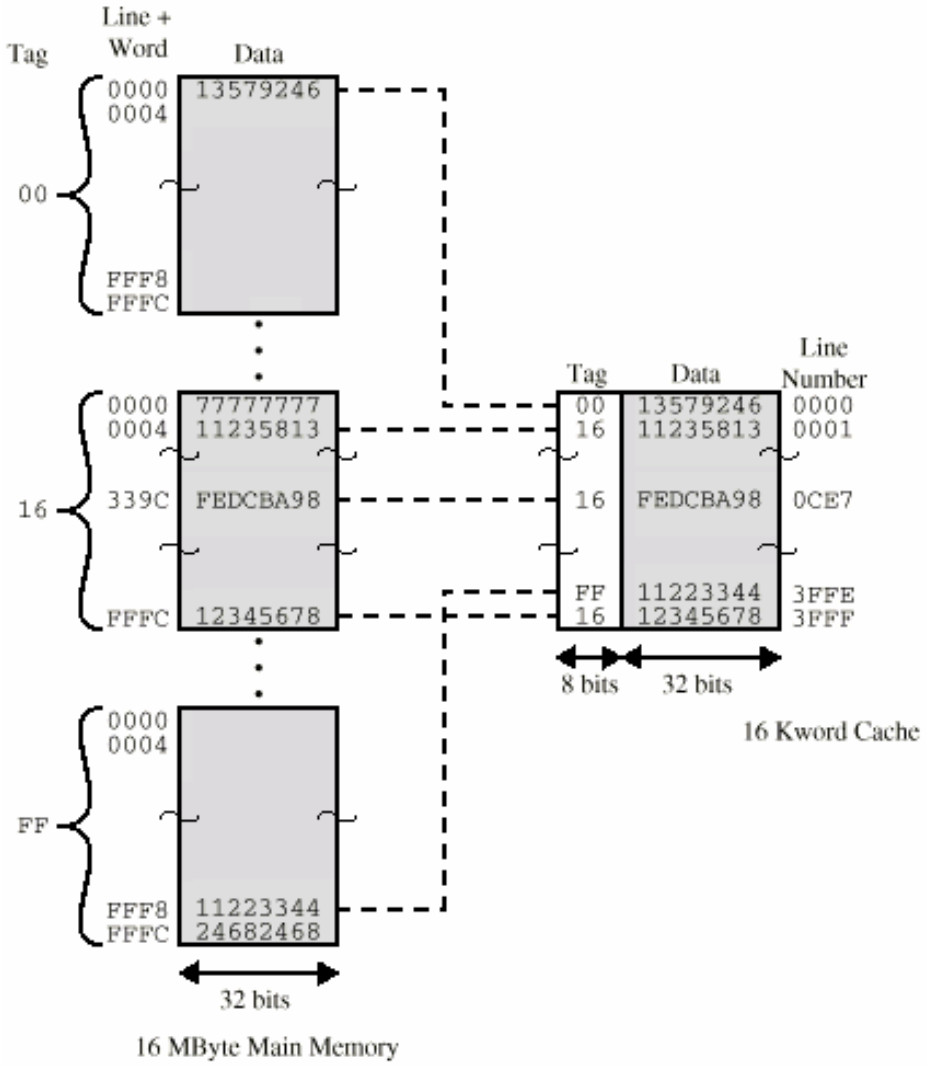
- Línea del Cache — Bloques de Memoria

Línea del Cache	Bloques de Memoria
0	0, m, 2m, 3m...2^s-m
1	1,m+1, 2m+1...2^s-m+1
m-1	m-1, 2m-1,3m-1...2^s-1

Organización Cache de Mapeo Directo



Ejemplo de Mapeo Directo



Resumen de Mapeo Directo

- Longitud dirección = (s + w) bits
- Numero de unidades direccionables = 2^{s+w} palabras o bytes
- Tamaño del bloque = tamaño de línea = 2^{w} palabras o bytes
- Numero de bloques en memoria principal = $2^{s+ w}/2^{w} = 2^{s}$
- Número de líneas del cache = m = 2^{r}
- Tamaño de la etiqueta = (s – r) bits

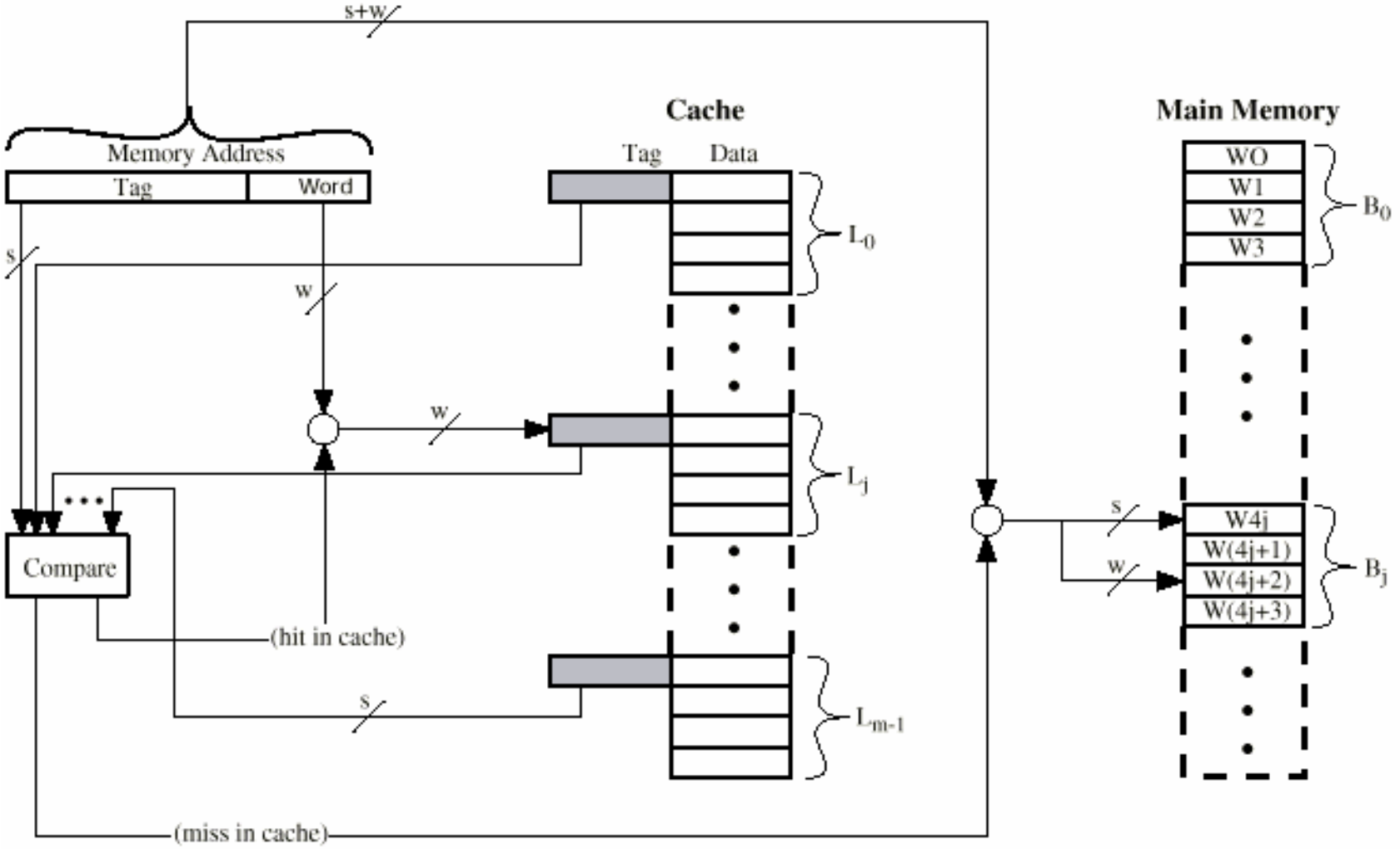
Mapeo Directo: Ventajas y Desventajas

- Simple
- Económico
- Posición fija de cache para cada bloque
  - Si un programa accede a 2 bloques que están mapeados en la misma línea repetidamente, las fallas de cache aumentan

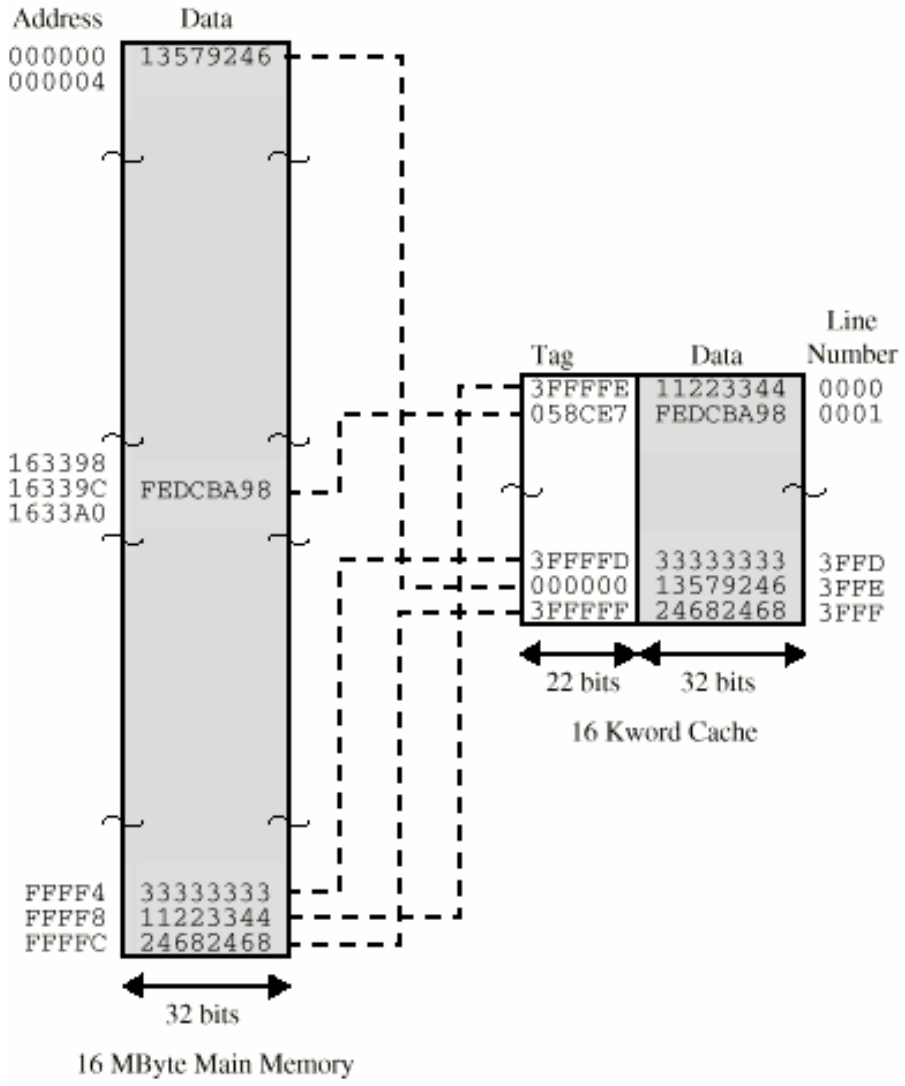
Mapeo Asociativo

- Cada bloque de memoria puede almacenarse en cualquier línea del cache
- La dirección es interpretada como etiqueta y palabra
- La etiqueta identifica el bloque
- Se busca en todas las etiquetas para verificar
- Esta búsqueda consume tiempo

Organización de Cache Asociativa Total



Ejemplo de Mapeo Asociativo



Estructura de direccionamiento del Mapeo Asociativo

Tag 22 bit	Word 2 bit

- Etiqueta de 22 bit se almacena con los cada uno de los bloques de datos de 32 bit
- Se compara el campo etiqueta del cache con los 22 MSb para verificar la corespondencia
- Los 2 LSb de la dirección identifican la palabra en el bloque
- Ej.

Dirección	Etiqueta	Dato	Línea Cache
FFFFFC	3FFFFF	24682468	3FFF

Resumen del Mapeo Asociativo

- Longitud dirección = (s + w) bits
- Numero de unidades direccionables = 2^{s+w}
- Tamaño del bloque = tamaño de línea = 2^{w}
- Numero de bloques en memoria principal = $2^{s+w}/2^{w} = 2^{s}$
- Numero de líneas en cache = indeterminado
- Tamaño de la etiqueta = s bits

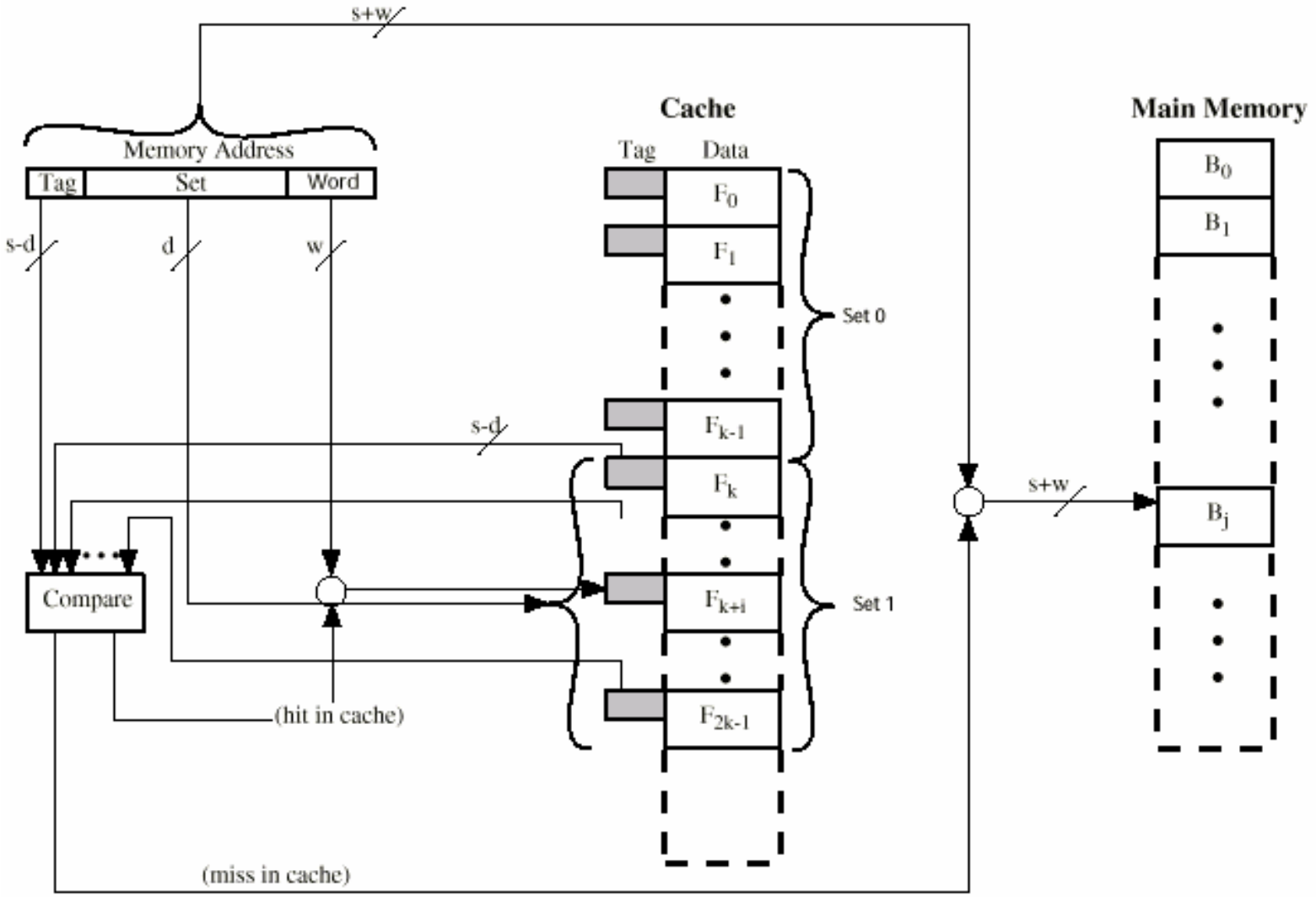
Mapeo Asociativo por Conjuntos

- Cache dividido en conjuntos (v)
- Cada conjunto contiene un numero de líneas (k)
- Un bloque se mapea en una línea de un conjunto
  - Ej. El bloque B puede estar en alguna línea del conjunto i
- Ej. 2 líneas por conjunto
  - mapeo asociativo de 2 vías
  - Un bloque puede estar en una de las dos líneas en un conjunto

Ejemplo de Mapeo Asociativo por Conjuntos

- Número de conjunto de 13 bits
- Numero de bloque en memoria es modulo 2^{13}
- 000000, 00A000, 00B000, 00C000 ... se mapean en el mismo conjunto

Organización de Cache Asociativo de Dos Vías



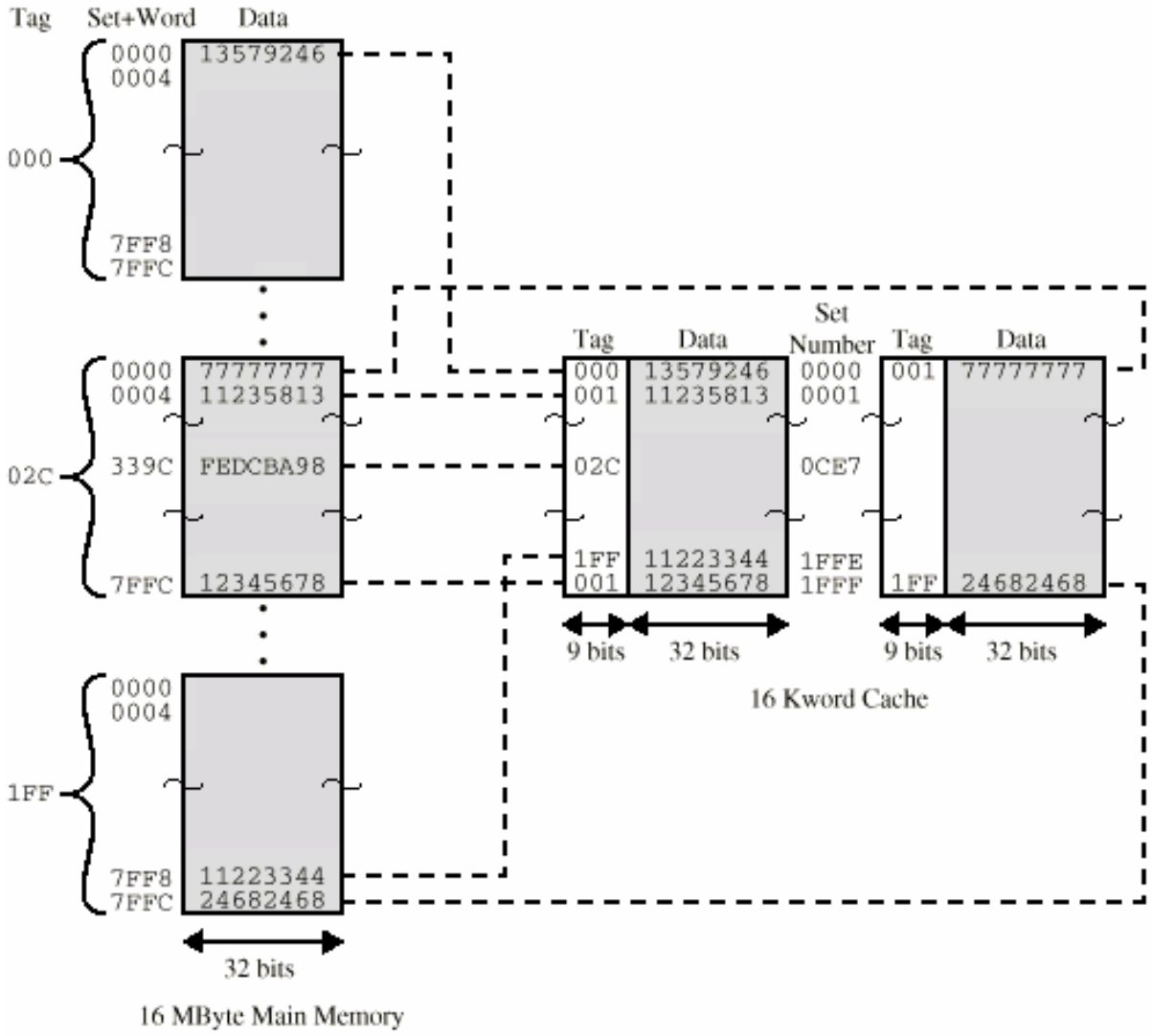
Estructura de direccionamiento del Mapeo Asociativo por conjuntos

Tag 9 bit	Set 13 bit	Word 2 bit

- Campo conjunto para determinar que conjunto mirar
- Campara el campo etiqueta para ver si hay coincidencia
- Ej.

Dirección	Eti.	Dato	Nº Conjunto
1FF 7FFC	1FF	12345678	1FFF
001 7FFC	001	11223344	1FFF

Ejemplo de Mapeo Asociativo por Conjuntos de 2 Vías



Resumen de Mapeo Asociativo por Conjuntos

- Longitud dirección = (s + w) bits
- Nº de unidades direccionables = 2^{s+w} words or bytes
- Tamaño del bloque = tamaño de línea = 2^{w}
- Nº de bloques en memoria = 2^{d}
- Nº de líneas por conjunto = k
- Nº de conjuntos = v = 2^{d}
- Nº de líneas en cache = kv = k * 2^{d}
- Tamaño de etiqueta = (s – d) bits

Algoritmos de Sustitución (1)
Mapeo Directo

- No hay elección
- Cada bloque solo se mapea en una línea
- Reemplaza la línea

Algoritmos de Sustitución (2) Asociativos

- Algoritmo implementado por hardware (velocidad)
- Least Recently used (LRU)-Menos recientemente usado
- Ej. En Asociativo por conjuntos de 2 vías
  - Cual de los 2 bloques es LRU?
- First in first out (FIFO)-1º en Entrar 1º en Salir
  - Reemplaza el bloque que ha estado en cache por mas tiempo
- Least frequently used - Menos frecuentemente usado
  - Reemplaza el bloque con menos aciertos
- Random - Aleatorio

Políticas de Escritura

- No debe sobrescribirse un bloque de cache si no se a actualizado a memoria
- Múltiples CPU pueden tener caches individuales
- E/S puede direccionar memoria directamente

Write through

- Las escrituras son tanto en cache como en memoria
- Múltiples CPU pueden vigilar el trafico en memoria para mantener actualizado su cache
- Mucho tráfico
- Escrituras hacen lento el bus

Write back

- Se actualiza inicialmente solo en cache
- Bit de actualización para líneas vigilar su coherencia con memoria
- Si un bloque es reemplazado en cache solo si el bit esta activado
- Otros caches interfieren con el sincronismo
- E/S deben acceder a memoria a través del cache
- 15% de referencias a memoria son escritura

Tamaño de Bloque

- Aumenta el tamaño y aumentaría la tasa de aciertos
- Para bloques grandes, disminuye la cantidad de los mismos en cache
- También disminuye la probabilidad encontrar la referencia deseada
- 4 a 8 unidades direccionables

Número de Caches

- Uno o mas Niveles
  - On Chip (libera el bus) (L1)
  - Externa (SDRAM) (L2)
- Unificado o Dividido
  - Unificada
    - Simple
    - Se adecua a la demanda
  - Datos e instrucciones
    - Elimina la competencia (fetch y ejecución)

Lecturas

- Capítulo 4 de Stallings