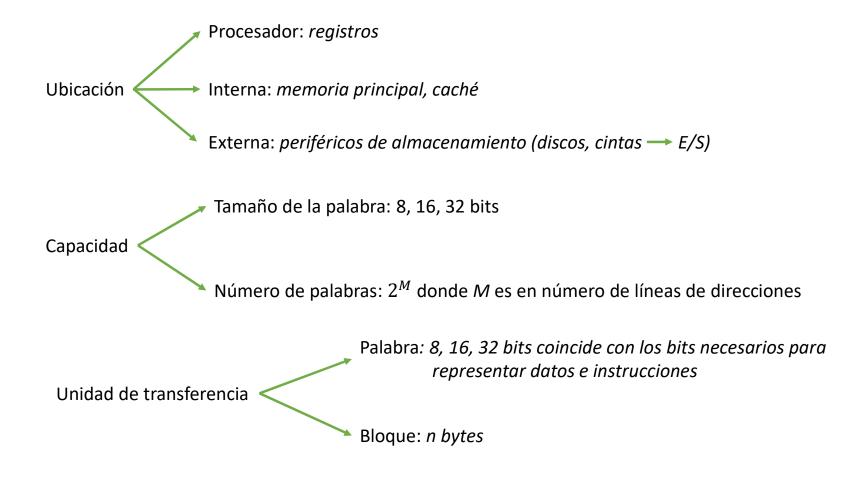
# **MEMORIA CACHÉ**



Ninguna tecnología es óptima para satisfacer las necesidades de un computador

Jerarquía de subsistemas de memoria

## Características de los sistemas de memoria



Secuencial: *unidades de datos (registros) – secuencia lineal específica* ej.: unidades de cinta. Mecanismo R/WR.

Método de acceso

Directo: mecanismo R/WR se posiciona directamente en la vecindad del bloque o registro y luego prosigue de forma secuencial. ej.: unidades de disco

Aleatorio: cada posición de memoria tiene un único mecanismo de direccionamiento cableado

Asociativo: tipo aleatorio. Permite comparar simultáneamente, para todas las palabras, un patrón de bits. Una palabra es recuperada en función de su contenido en lugar de su dirección. Ej.: memoria caché

Prestaciones

Tiempo de acceso: tiempo necesario para completar una operación de RD/WR

Tiempo de ciclo: *Tiempo de acceso + tiempo adicional.*Depende de las características del bus

Velocidad de transferencia: velocidad de transferencia de datos a, o desde, una unidad de memoria

Para memorias de acceso aleatorio = tiempo de ciclo

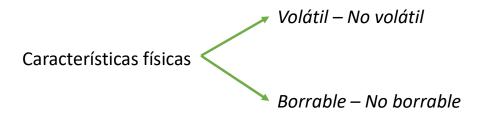
Para otras memorias  $T_N = T_A + \frac{N}{R}$ 

 $T_N$  es el tiempo medio de RD/WR de N bits

 $T_A$  es el tiempo de acceso medio

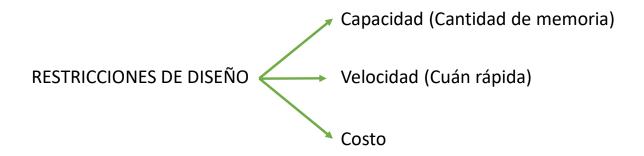
N es el número de bits

R es la velocidad de transferencia en bits por segundo



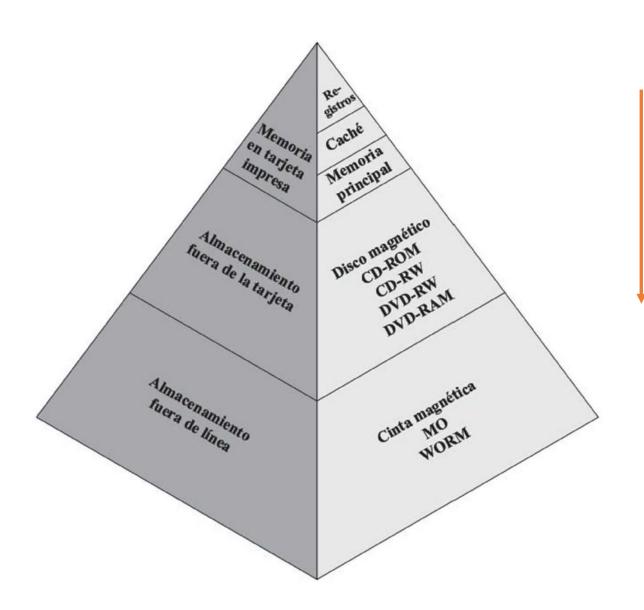
Organización ———— RAM: estructura física en bits para formar palabras

# JERARQUÍA DE MEMORIA



En el espectro de las tecnologías se cumplen las siguientes relaciones

A menor tiempo de acceso, mayor costo por bit
 A mayor capacidad, menor costo por bit
 A mayor capacidad, mayor tiempo de acceso



Disminuye el costo por bit

Aumenta la capacidad

Aumenta el tiempo de acceso

Disminuye la frecuencia de acceso a la memoria por parte del procesador

# Principio de localidad de las referencias

Base de la mejora de las prestaciones de un sistema con varios niveles de memoria

Excepto por las instrucciones de bifurcación o llamadas, un programa se ejecuta en forma secuencial

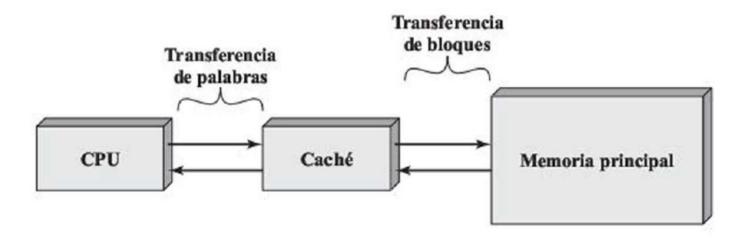
Un programa queda confinado en una ventana estrecha de nivel de anidamiento de procedimientos

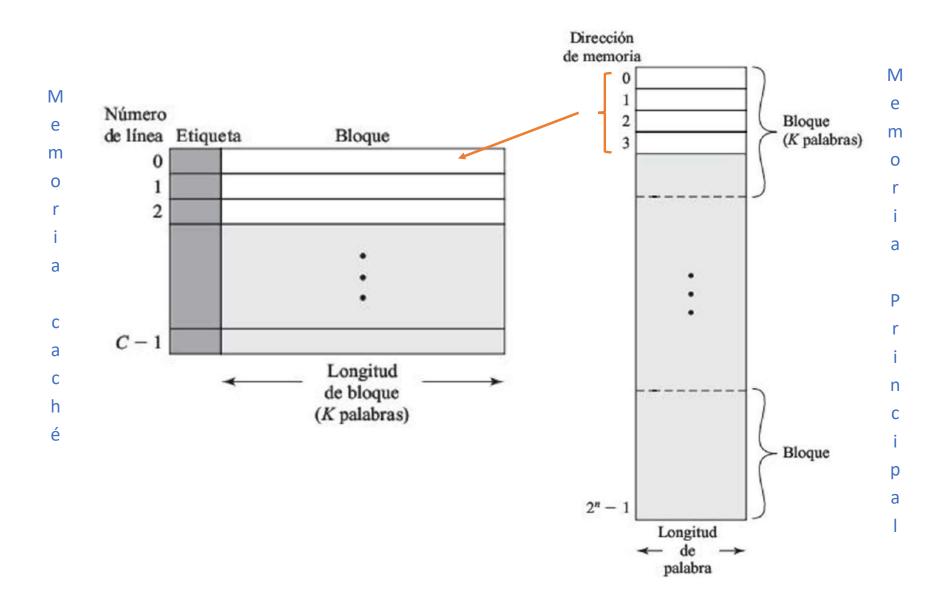
La mayoría de las construcciones iterativas consta de un número pequeño de instrucciones

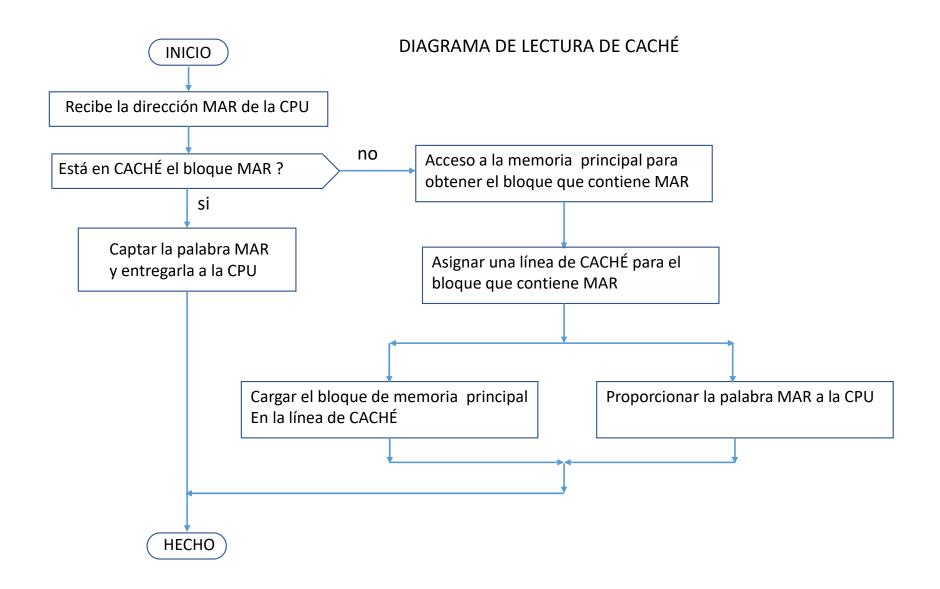
Las referencias sucesivas a estructura de datos serán a unidades de datos próximos entre si

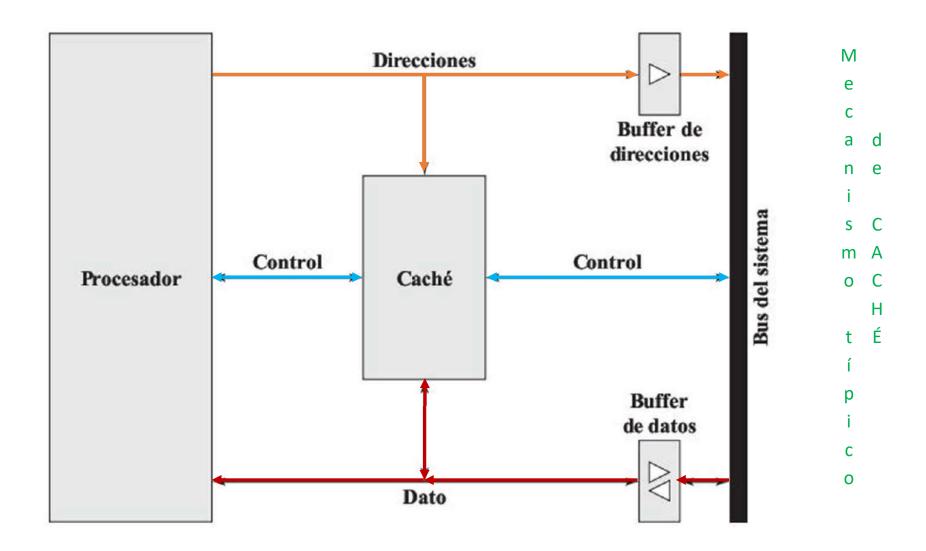
El programa se ejecuta por Clusters

# PRINCIPOS BÁSICOS DE LA MEMORIA CACHÉ









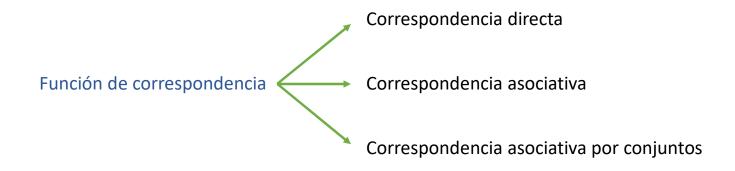
## ELEMENTOS DE DISEÑO DE UNA CACHÉ

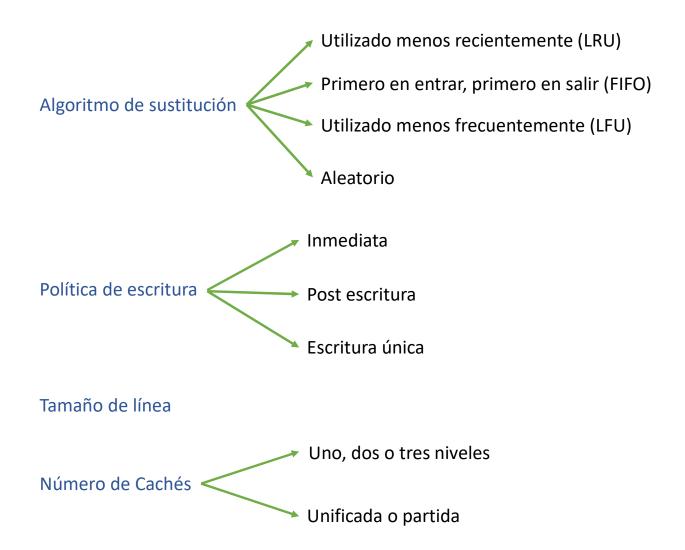
#### Tamaño de la Caché

la prestación de la Caché es muy sensible al tipo de tarea.

Es imposible predecir el tamaño óptimo

Suficientemente pequeño para que el costo medio por bit se aproxime al de la memoria principal Suficientemente grande para que el tiempo de acceso medio total se aproxime al de la Caché sola





# Función de correspondencia

Memoria principal organizada en bytes con bus de direcciones de 24 bits :  $2^{24} = 16 MBytes$ 

Memoria Caché de 64KBytes

Tamaño de bloque de la memoria principal = tamaño de línea de la Caché = 4bytes

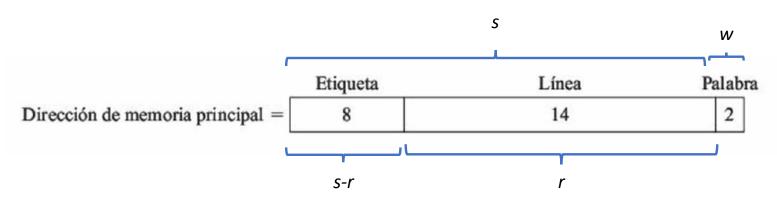
La cantidad de bloques de la memoria principal es 16MBytes / 4Bytes = 4MBloques =  $2^{22}$ Bloques

La cantidad de líneas de la memoria Caché es 64KBytes / 4Bytes = 16Klíneas =  $2^{14}$  Líneas

# Correspondencia directa

A cada bloque de la memoria principal le corresponde solo una línea posible de caché

#### Dirección de memoria de 24 bits



Longitud de las direcciones: s + w bits

Número de unidades direccionables:  $2^{s+w}$  bytes

Tamaño de bloque = tamaño de línea:  $2^w$  bytes

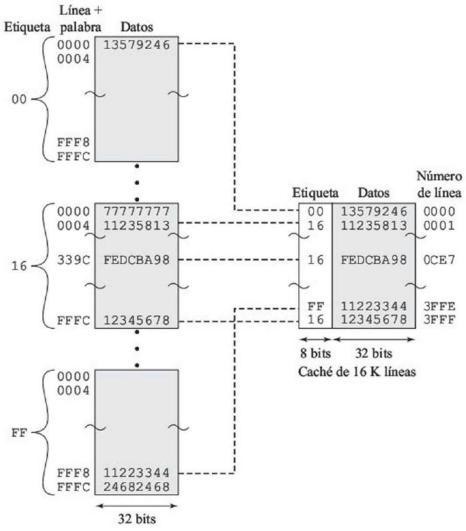
Número de bloques en la memoria principal:  $\frac{2^{s+w}}{2^w} = 2^s$ 

Número de líneas en la caché:  $2^r = m$ 

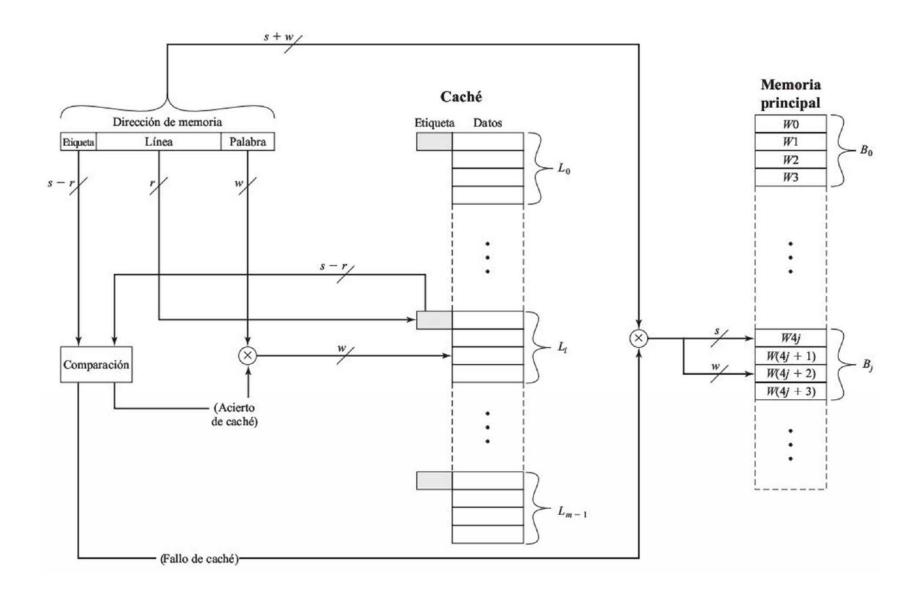
Tamaño de la etiqueta: s - r bits

## Inconveniente

 $a \ 2^{s-r}$  bloques de la memoria principal le corresponde la misma línea de la caché



Memoria principal de 16 MB



# Correspondencia asociativa

Cada bloque de la memoria principal puede cargarse en cualquier línea de la memoria caché

	S	W
	Etiqueta	Palabra
Dirección de memoria principal =	22	2

Longitud de las direcciones: s + w bits

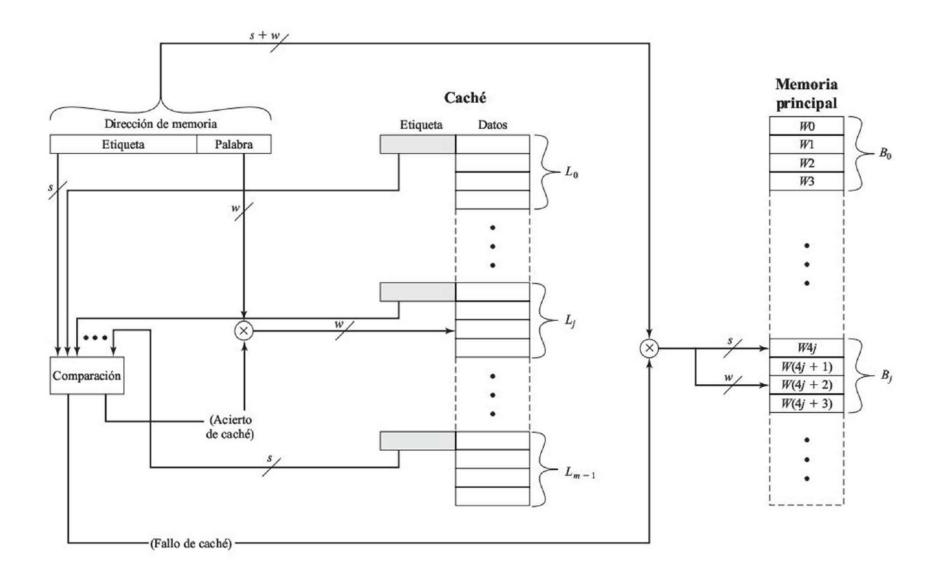
Número de unidades direccionables:  $2^{s+w}$  bytes

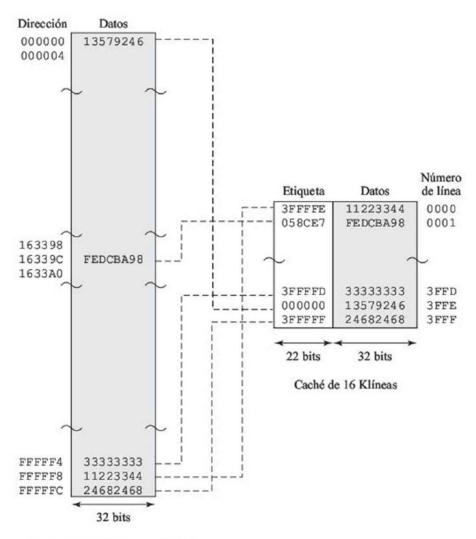
Tamaño de bloque = tamaño de línea:  $2^w$  bytes

Número de bloques en la memoria principal:  $\frac{2^{s+w}}{2^w} = 2^s$ 

Número de líneas en la caché: indeterminado

Tamaño de la etiqueta: s bits

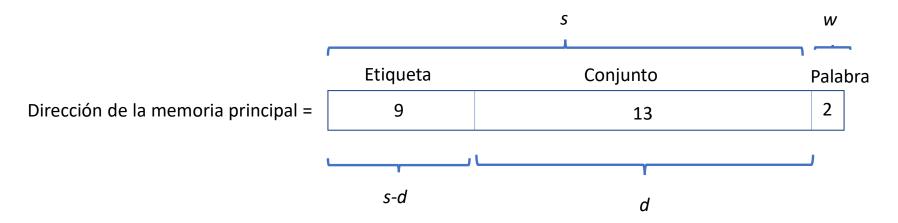




Memoria principal de 16 MBytes

## Correspondencia asociativa por conjuntos

La caché se divide en *v* conjuntos de *k* líneas (vías) cada uno



Longitud de las direcciones: s + w bits

Número de unidades direccionables:  $2^{s+w}$  bytes

Número de conjuntos:  $v = 2^d$ 

Tamaño de la etiqueta: *s - d bits* 

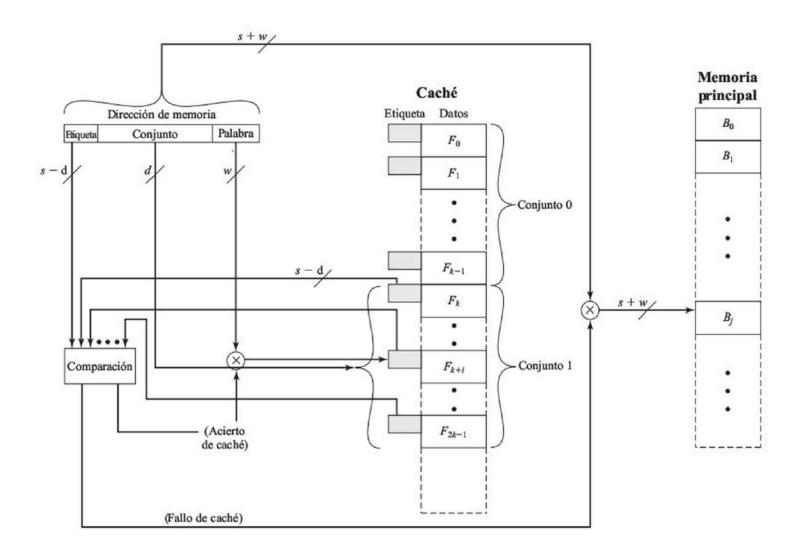
Tamaño de bloque = tamaño de línea:  $2^w$  bytes

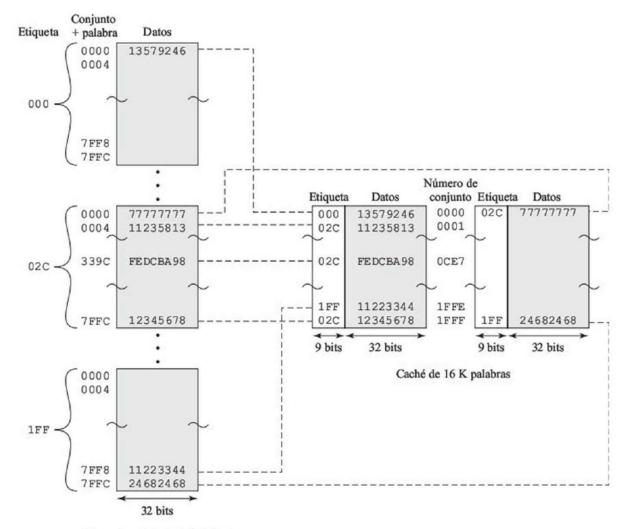
Número de líneas en el conjunto: k

Número de líneas en la caché:  $k \cdot v = k \ 2^d$ 

Caso típico: k = 2

Aumenta significativamente el número de aciertos





Memoria principal de 16 Mbytes

# Algoritmos de sustitución

Correspondencia directa: Hay una sola posible línea para cada bloque particular. No hay elección posible Thrashing

Para las técnicas asociativas:

Utilizado menos recientemente (LRU):

Sustituye el bloque que ha permanecido mas tiempo en la caché sin ser referenciado

Primero en entrar, primero en salir (FIFO)

Se sustituye el bloque que ha permanecido mas tiempo en la caché

Utilizado menos frecuentemente (LFU)

Sustituye el bloque menos referenciado

Aleatorio

Se sustituye una línea al azar

#### Política de escritura

Contenido de la memoria principal = Contenido de la memoria caché

Escritura inmediata: La escritura se hace tanto en cache como en memoria principal

Coherencia de caché

Varios procesadores con su propia caché —— Monitorear el tráfico con memoria principal

Incremento sustancial del tráfico con memoria principal ——— Cuello de botella

Post escritura 

Minimiza las escrituras 

Se realizan solo en caché

Reemplazo una línea que fue modificada ——— Actualiza el bloque correspondiente de la memoria principal

Porciones de memoria no válida

Los accesos de los módulos de E/S deben hacerse a través de la caché

Varios procesadores con su propia caché —— Coherencia de caché

Vigilancia bus con escritura inmediata

Transparencia hardware

Memoria excluida de caché

#### Tamaño de línea

Principio de localidad —— Aumento del tamaño de bloque —— Aumento de aciertos

Bloques mas grandes disminuye el número de bloques que caben en la caché

Bloque mas grande —— Palabra adicional mas lejana —— Disminuye la probabilidad a corto plazo

## Número de cachés

Caches multinivel Niveles de caché L1, L2 y L3 incluidas en el chip del procesador



Competición por la cache