Arquitectura de Computadoras

UNTDF

Profesor Guillermo Eduardo Prisching

E-mail: gprisching@untdf.edu.ar

Memorias

Cómo potenciar el rendimiento:

- **Mejorar la implementación de hardware**
- **Mejorar la arquitectura de hardware**

D+I de técnicas para aumentar el rendimiento

1) Memoria Caché

2) Predicción de Ramas

3) Ejecución fuera de orden c/cambio de registros

4) Ejecución especulativa

Memoria Caché

Objetivo: proporcionar instrucciones o datos de Memoria Principal tan rápido como sea posible

1) Bajar la latencia

2) Aumentar el ancho de banda

Memoria Caché

Primera aproximación

Utilizar una MC para los datos y otra MC para las instrucciones a.k.a “Caché Dividida”

Se requiere una arquitectura que brinde acceso independiente a MP a cada una de las MCs

Memoria Caché

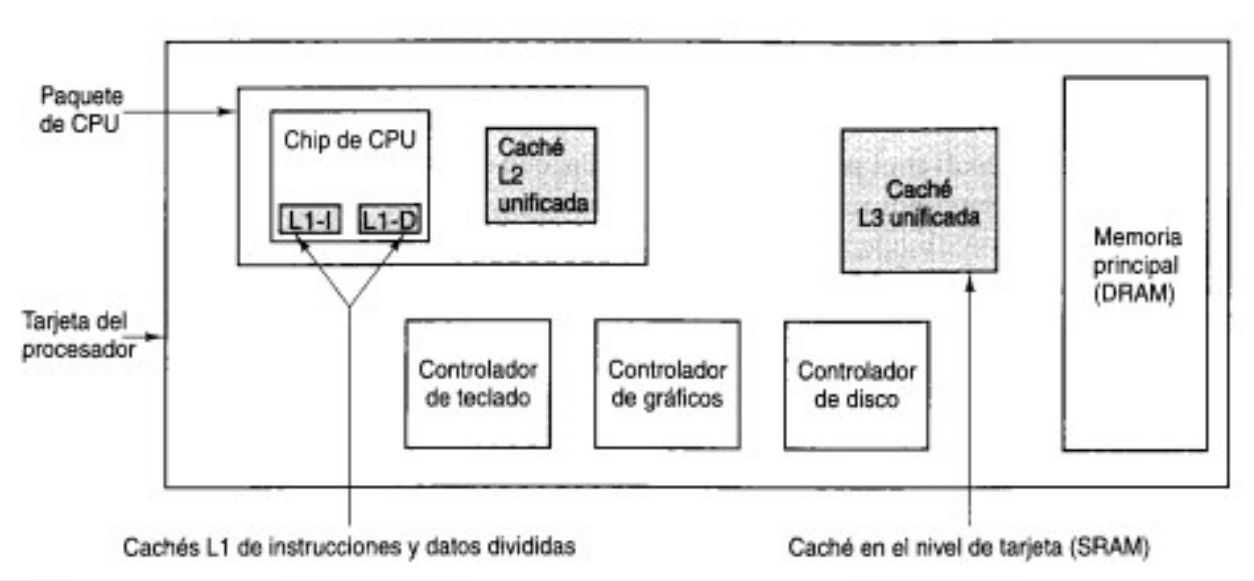
Segunda aproximación

Utilizar varios niveles de MC

Se agregan diferentes niveles de Caché entre el Procesador y la MP

Trade off → a mayor capacidad menor velocidad de acceso

Memoria Caché



Sistema con tres niveles de Caché (A.Tanenbaum)

Memoria Caché

- **Propiedad Inclusiva: todos los niveles están relacionados en conjuntos y sub-conjuntos**
- **Propiedad de Localidad de referencia: temporal, espacial**

Memoria Caché

- **Se organizan en bloques de tamaño fijo denominados línea o entrada o renglón**
- **Tamaño típico de cada línea va de 4Bytes a 64KBytes**
- **Cada línea se numera secuencialmente comenzando desde el 0**

Memoria Caché

- **Ante una referencia a MP el circ. controlador verifica si está en la MC**
- **Si no, se elimina la línea de MC y se busca en MP o en una Caché de nivel inferior.**
- **Una vez recuperada se reemplaza en la última MC accedida**

Memoria Caché

- **Tasa de aciertos** → $h = k - 1 / k$
- **Tasa de fallos** → $1 - h$
- **Tiempo de acceso medio** → $c + (1 - h) m$

Una Memoria Caché mal diseñada (baja tasa de aciertos) es peor que no tenerla

Memoria Caché

Función de correspondencia

- **Dado que la MC es de inferior capacidad a la MP debe elegirse como almacenar una porción de ésta**

- **Existen tres aproximaciones**
  1. **Mapeo Directo**
  2. **Mapeo Asociativo**
  3. **Mapeo Asociativo por Conjuntos**

Memoria Caché

Correspondencia Directa

¿porqué se llama Mapeo Directo?

Porque cada bloque de memoria tiene asignado ***uno y solo un renglón/entrada*** *de la Memoria Caché*

Memoria Caché

Descripción de una línea/entrada/renglón de Caché:

Bit de validez: indica si la entrada contiene datos válidas. Cuando arranca el sistema todas las entradas son NO VÁLIDAS.

Campo Etiqueta (Tag): identifica al bloque de memoria de donde provienen los datos

Campo de Datos: contiene una copia de los datos de Memoria Principal.

Memoria Caché

Supongamos el siguiente escenario:

MC: 64KBytes
Long de línea: 32Bytes

MP: 4MBytes
Long de palabra de MP: 4Bytes (32bits)

Memoria Caché

Entrada	Bit de validez	Etiqueta	Datos	Direcciones de M.Ppal que usan esta entrada
0	x			0-7, 16384-16391, 32768-32775
1	x			8-15, 16392-16399,32776-32783
2	x			16-23, 16400-16407, 32784-32791
*	*	****	****	****
*	*	****	****	****
*	*	****	****	****
2047	x			16376-16383, 32760-32767, ...

Memoria Caché

Función de correspondencia (*Método de acceso a la MC*)

Se utiliza la dirección de MP dividiendola en tres campos:

(1)	(2)	(3)
Etiqueta (Tag)	#Línea	Palabra

(1) Corresponde a los bits almacenados como TAG en una línea/entrada de la Caché

(2) Indica cuál línea entrada de Caché contiene los datos correspondientes, si están presentes.

(3) Indica a cuál palabra se hace referencia dentro de la línea/entradas

Memoria Caché

Expresión de correspondencia DIRECTA:

$$i = j \bmod m$$

i : nro línea de la MC
j: nro de bloque de la MP
m: cantidad de líneas en la MC

Memoria Caché

Dirección de la MP = $s + w$

Etiqueta (Tag)	Linea	Palabra
s-r	*r*	*w*

w son los bits menos significativos e identifican cada palabra dentro de un bloque de memoria principal.

s son los bits restantes de la palabra de dirección de Memoria Principal y especifican uno de los 2^s bloques de la Memoria Ppal.

La lógica de la caché interpreta estos ***s*** bits como una Etiqueta Taag de ***s – r*** bits (parte más significativa)

r es el campo de línea el cuál identifica a una de las $m = 2^r$ líneas de la Caché.

Memoria Caché

Conclusiones

*La técnica de **Mapeo Directo** es simple y poco costosa de implementar. Su principal desventaja es que hay una posición concreta de Caché para cada bloque dado.*

Favorece intercambios continuos con la MP bajando la tasa de aciertos h

Memoria Caché

Supongamos el siguiente escenario:

MC: 64KBytes
Long de línea: 32Bytes
Líneas de MC: 2048

MP: 4MBytes
Long de palabra de MP: 4Bytes (32bits)
Lineas de direcciones: 20

Arquitectura de Computadoras

UNTDF

MP 4MB

Addr	Datos	Bloque
0		B0
***	***	
7		
8		B1
***	***	
15		
***	***	***
16376		B2047 *0*
***	***	
16383		
16384		B0
***	***	
16391		
***	***	***
32760		B2047 *1*
***	***	
32767		
1,048,575		B2047 *128*
1,048,575		

Linea	Tag	Campo Datos 32Bytes == 8 Palabras MP							
0									
2047									

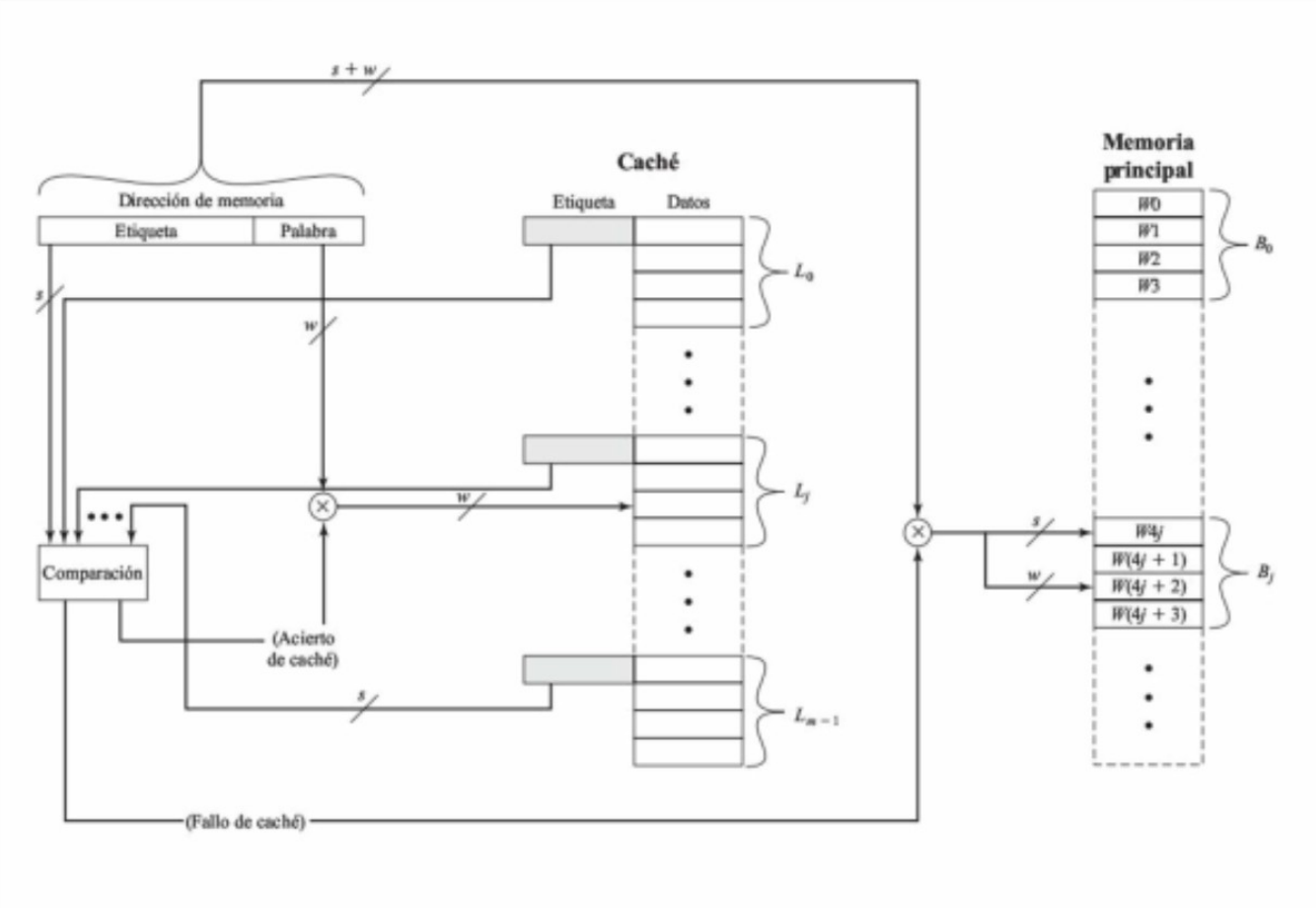
Memoria Caché

Correspondencia ASOCIATIVA:

- **Resuelve la desventaja de la correspondencia DIRECTA**
- **Este método permite que cada bloque de MP pueda cargarse en cualquier línea de la caché.**
- **Para lograrlo, el campo de etiqueta identifica UNIVOCAMENTE un bloque de MP**

Arquitectura de Computadoras

UNTDF



Organización de MC asociativa – Williams Stallings

Memoria Caché

Correspondencia ASOCIATIVA:

- **Una dirección de memoria se divide en dos campos 1) para la etiqueta (Tag), 2) para la palabra dentro de la línea de la MC**
- **Mayor flexibilidad → Cualquier bloque puede ser reemplazado cuando es necesario escribir en la MC**
- **Desventaja: requiere electrónica + compleja para examinar en paralelo las todas los Tag**

Memoria Caché

Correspondencia Asociativa

Determinar como serán las etiquetas y campo de palabra en el siguiente escenario:

MC: 64KBytes
Long de línea: 32Bytes
Líneas de MC: 2048

MP: 4MBytes
Long de palabra de MP: 4Bytes (32bits)
Lineas de direcciones: 20

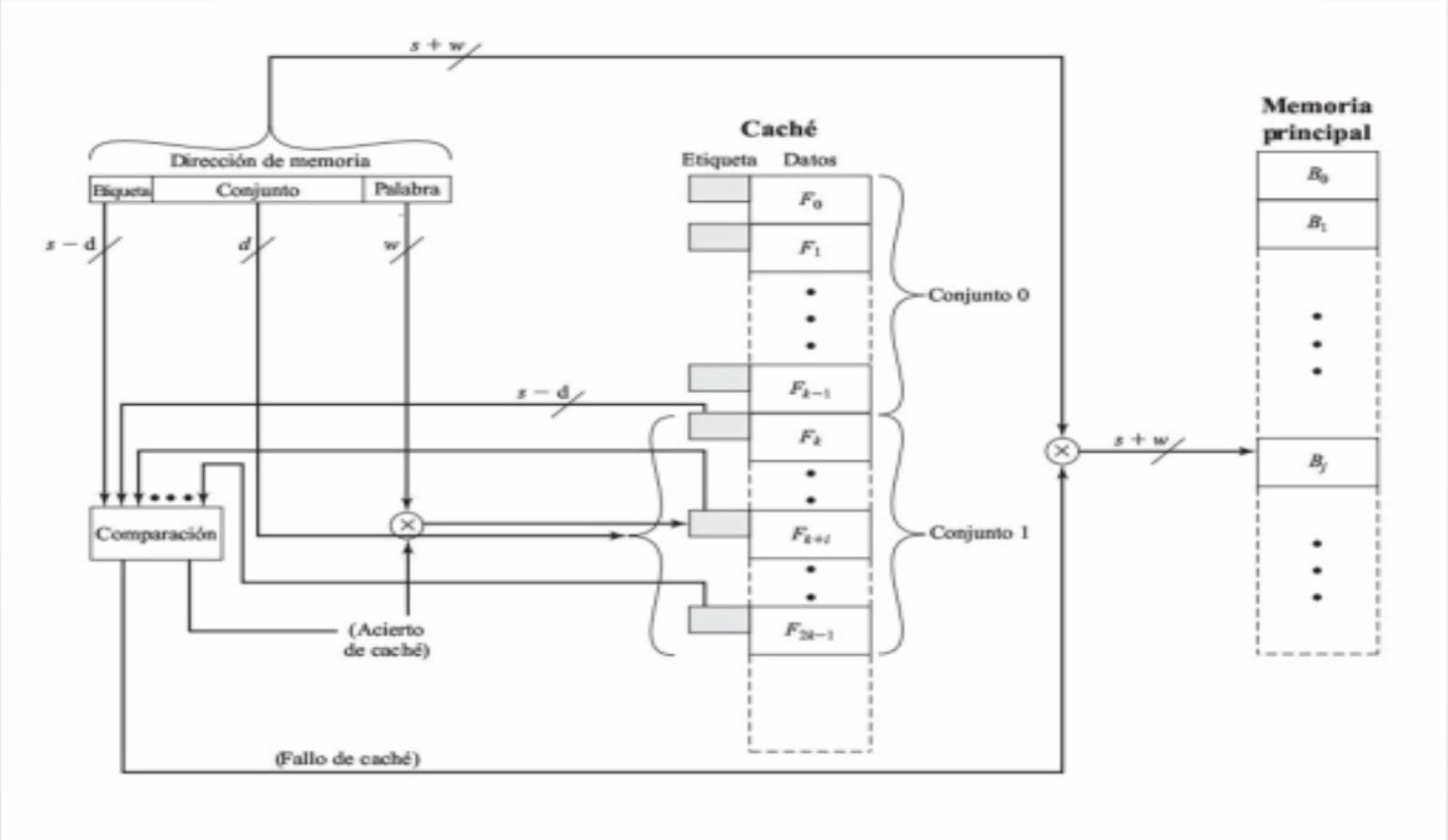
Memoria Caché

Correspondencia ASOCIATIVA POR CONJUNTOS:

- **Converge lo mejor de los dos mundos**
- **Un bloque B_j puede asignarse a cualquiera de las *k* líneas del conjunto *i* de la MC**
- **La etiqueta de de una dirección de MP es mucho más corta y se compara sólo con las etiquetas del mismo conjunto**

Arquitectura de Computadoras

UNTDF



Organización de MC asociativa por conjuntos – William Stallings

Memoria Caché

Correspondencia ASOCIATIVA POR CONJUNTOS:

$$m = v \; x \; k$$
$$i = j \; mod \; v$$

i = número de conjunto de cache
j = número de bloque de memoria principal
m = número de líneas de la cache
v = cantidad de conjuntos
k = cantidad de líneas del conjunto

Memoria Caché

Correspondencia ASOCIATIVA POR CONJUNTOS:

Etiqueta (tag)	Línea	Palabra

Dirección de memoria principal

Memoria Caché

Correspondencia ASOCIATIVA POR CONJUNTOS:

Existe dos casos extremos

1) $v = m$
2) $v = 1$

Memoria Caché

Correspondencia ASOCIATIVA POR CONJUNTOS:

Existe dos casos extremos

1) $v = m, k = 1 \rightarrow$ *se comporta como una correspondencia directa*
2) $v = 1, k = m \rightarrow$ *se comporta como una asociativa completa*