# Memoria Cache Organización del Computador I

#### David Alejandro González Márquez

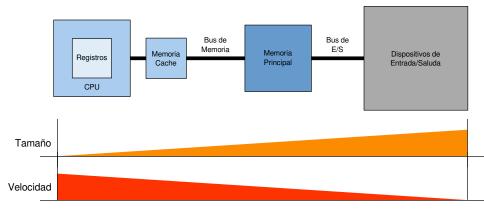
Departamento de Computación Facultad de Ciencias Exactas y Naturales Universidad de Buenos Aires

27.02.2018

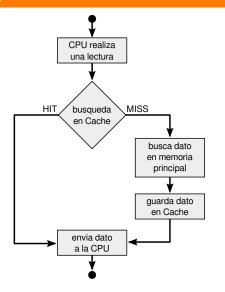
# Agenda

- Introducción
- características
- Tipos de Cache
- Ejercicios

# Jerarquía de Memoria



#### ¿Cómo funciona la Memoria Cache?



#### Eventos tras un acceso:

- Hit: El dato solicitado se encuentra en caché
- Miss: En caso contrario

#### Métricas:

- **Hit Rate** =  $\frac{\#hits}{\#pedidos}$
- Miss Rate =  $\frac{\#miss}{\#pedidos}$

#### Localidad Espacial y Localidad Temporal

#### Localidad Espacial:

Si se pide un dato en memoria, es altamente probable que a continuación se pida también otro dato que sea próximo en memoria al primero. ej: ejecución secuencial, recorrido de arrays, etc.

#### Localidad Temporal:

Si se pide un dato de memoria, es altamente probable que éste vuelva a ser reutilizado en un futuro inmediato. ej: variables, ciclos, etc.

Durante un *miss*, la memoria caché solicita a la memoria principal el dato buscado junto con los datos vecinos, de esta forma anticipa los futuros pedidos.

### Tipos de Cache

#### ■ Totalmente Asociativa:

Cada bloque en cache puede contener cualquier dato de la memoria principal

#### Correspondencia Directa:

Los bloques de cache almacenan direcciones específicas de memoria principal

#### Asociativa por Conjuntos:

Los bloques de cache se dividen en conjuntos y cada uno puede almacenar un conjunto de direcciones específicas de memoria principal

#### Política de reemplazo

- First In First Out (FIFO): El primer dato en entrar es el primero en descartado
- Least Recently Used (LRU):
  Se descarta el dato menos recientemente usado
- Least Frequently Used (LFU):
  Se descarta el bloque menos frecuentemente usado

#### Política de Escritura

#### Write-through:

Se escribe en cache y en memoria al mismo tiempo

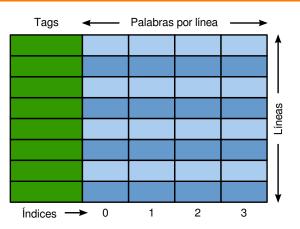
#### ■ Write-back:

Se escribe en memoria principal cuando se desaloja el dato de cache

### Tipos de fallos

- Forzosos (Compulsory):
   En el primer acceso éste no se encuentra en la caché (primera referencia)
- Capacidad (Capacity):
   La caché no puede contener todos los bloques necesarios durante la ejecución de un programa
- Conflicto (Conflict):
   Diferentes bloques deben ir necesariamente al mismo conjunto o línea (fallos de colisión)

#### Estructura de una Caché



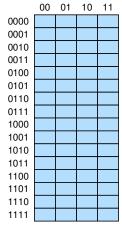
- Línea: Mínima unidad de almacenamiento la memoria cache
- Índice: Indica la posición del dato dentro de una línea
- Tag: Identificador del dato en memoria principal

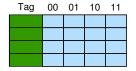
# Ejemplos de memorias cache

- Considerar una memoria de 64 unidades direccionables
- y una memoria cache de 16 unidades direccionables
- con las siguientes características:
  - Totalmente Asociativa: 4 líneas, 4 unidades por línea
  - Correspondencia Directa: 4 líneas, 4 unidades por línea
  - Asociativa por Conjuntos: 2 conjuntos o vías, 2 líneas por conjunto, 4 unidades por línea

#### Totalmente Asociativa









#### Totalmente Asociativa

#### Memoria Principal

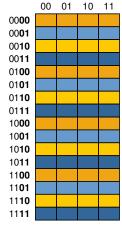
85 75 65 52
65
52
62
39
42
55
69
34
16
44
43
41
34
33

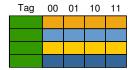
Tag	00	01	10	11
0100	21	43	44	62
0110	59	11	21	42
0101	60	52	70	39
0001	51	64	45	75

Tag	Indice	
4	2	6 bits

## Correspondencia Directa

#### Memoria Principal





Tag	Línea	Línea Indice	
2	2	2	6 bits

# Correspondencia Directa

#### Memoria Principal

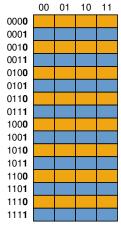
	00	01	10	11
<b>0000</b>	62	83	33	85
00 <b>01</b>	51	64	45	75
00 <b>10</b>	42	88	42	65
00 <b>11</b>	82	81	33	52
01 <b>00</b>	21	43	44	62
01 <b>01</b>	60	52	70	39
01 <b>10</b>	59	11	21	42
01 <b>11</b>	32	81	54	55
10 <b>00</b>	45	57	53	69
10 <b>01</b>	63	13	43	34
10 <b>10</b>	98	29	84	16
10 <b>11</b>	64	77	55	44
11 <b>00</b>	72	62	37	43
11 <b>01</b>	84	41	68	41
11 <b>10</b>	77	86	77	34
11 <b>11</b>	49	99	51	33

Tag	00	01	10	11
11	72	62	37	43
10	63	13	43	34
01	59	11	21	42
11	49	99	51	33

Tag	Línea Indice		
2	2	2	6 bits

## Asociativa por Conjuntos

#### Memoria Principal





Tag	Línea	Indice	
3	1	2	6 bits

# Asociativa por Conjuntos

#### Memoria Principal

00	υı	10	- 1 1
62	83	33	85
51	64	45	75
42	88	42	65
82	81	33	52
21	43	44	62
60	52	70	39
59	11	21	42
32	81	54	55
45	57	53	69
63	13	43	34
98	29	84	16
64	77	55	44
72	62	37	43
84	41	68	41
77	86	77	34
49	99	51	33
	62 51 42 82 21 60 59 32 45 63 98 64 72 84	62 83 51 64 42 88 82 81 21 43 60 52 59 11 32 81 45 57 63 13 98 29 64 77 72 62 84 41 77 86	62 83 33 51 64 45 42 88 42 82 81 33 21 43 44 60 52 70 59 11 21 32 81 54 45 57 53 63 13 43 98 29 84 64 77 55 72 62 37 84 41 68 77 86 77

Tag	00	01	10	11	Tag	00	01	10	11
					000				
011	32	81	54	55	001	82	81	33	52

rag	Linea	Indice	
3	1	2	6 bits

**Memoria Principal** 2<sup>20</sup> bytes, direccionable a byte. **Cache** 32 líneas de 16 bytes cada una.

¿Cuánto mide un bloque para esta configuración?

- ¿Cuánto mide un bloque para esta configuración?
- ¿Cuántos bloques entran en memoria principal?

- ¿Cuánto mide un bloque para esta configuración?
- ¿Cuántos bloques entran en memoria principal?
- ¿Cuántas líneas entran en un bloque?

- ¿Cuánto mide un bloque para esta configuración?
- ¿Cuántos bloques entran en memoria principal?
- ¿Cuántas líneas entran en un bloque?
- ¿Cuántas direcciones distintas refieren a una misma línea?

- L'Cuánto mide un bloque para esta configuración?
- ¿Cuántos bloques entran en memoria principal?
- ¿Cuántas líneas entran en un bloque?
- ¿Cuántas direcciones distintas refieren a una misma línea?
- ¿Cómo puedo saber si está cargada la línea donde se encuentra la palabra referida por la dirección C34A6?

¿Cuánto mide un bloque para esta configuración? Para esta configuración, coincide con el tamaño de la cache

$$\#$$
 líneas cache x capacidad de una línea  $=32\frac{\text{líneas}}{\text{cache}} \times 16\frac{B}{\text{línea}} = 512\frac{B}{\text{cache}} = 2^9\frac{B}{\text{cache}}$ 

¿Cuántos bloques entran en memoria principal?

$$\frac{\text{capacidad memoria}}{\text{tama\~no bloque}} = \frac{2^{20}B/MP}{2^{9}B/bloque} = \frac{2^{11}bloque}{MP}$$

¿Cuántas líneas entran en un bloque? Para esta configuración, tantas como entren en la cache:

$$\frac{\text{capacidad cache}}{\text{capacidad de una línea}} = \frac{512B/\textit{cache}}{16B/\textit{linea}} = \frac{2^9B/\textit{cache}}{2^4B/\textit{linea}} = 2^5\frac{\textit{linea}}{\textit{cache}}$$

¿Cuántas direcciones distintas refieren a una misma línea?

tamaño del tag x tamaño de la linea  $= 2^{11}$  bits  $\times 2^4$  bits  $= 2^{15}$  direcciones

#### ¿Cómo puedo saber si está cargada la línea donde se encuentra la palabra referida por la dirección C34A6?

Primero me fijo cuánto mide cada campo de una dirección de memoria para esta configuración de cache: **tag** 11 *bits*, **línea** 5 *bits*, **índice** 4 *bits*. Después, paso la dirección a binario para saber el valor de los campos correspondientes a esa dirección.

La dirección en binario:

$$\frac{C}{1100} \frac{3}{0011} \frac{4}{0100} \frac{A}{1010} \frac{6}{0110}$$

Agrupada según los campos tag, línea e índice:

$$\frac{61A}{110,0001,1010} \frac{A}{0,1010} \frac{6}{0110}$$

Finalmente, me tengo que fijar si en el lugar reservado para las líneas número 0xA, está cargada la línea correspondiente al bloque número 0x61A; si es así, la línea correspondiente a la dirección pedida está cargada en cache.

Memoria Principal 1 MB, direccionable a byte.

Cache 32 líneas de 64 bytes cada una, 2 vías.

¿Cuánto mide un bloque para esta configuración?

**Memoria Principal** 1 MB, direccionable a byte. **Cache** 32 líneas de 64 bytes cada una, 2 vías.

- ¿Cuánto mide un bloque para esta configuración?
- ¿Cuántos bloques entran en memoria principal?

**Memoria Principal** 1 MB, direccionable a byte. **Cache** 32 líneas de 64 bytes cada una, 2 vías.

- ¿Cuánto mide un bloque para esta configuración?
- ¿Cuántos bloques entran en memoria principal?
- ¿Cuántas líneas entran en un bloque?

Memoria Principal 1 MB, direccionable a byte.

Cache 32 líneas de 64 bytes cada una, 2 vías.

- ¿Cuánto mide un bloque para esta configuración?
- ¿Cuántos bloques entran en memoria principal?
- ¿Cuántas líneas entran en un bloque?
- ¿Cuántas direcciones distintas refieren a una misma línea?

Memoria Principal 1 MB, direccionable a byte.

Cache 32 líneas de 64 bytes cada una, 2 vías.

- ¿Cuánto mide un bloque para esta configuración?
- ¿Cuántos bloques entran en memoria principal?
- ¿Cuántas líneas entran en un bloque?
- ¿Cuántas direcciones distintas refieren a una misma línea?
- ¿Cómo puedo saber si está cargada la línea donde se encuentra la palabra referida por la dirección C34A6?

¿Cuánto mide un bloque para esta configuración? Para esta configuración, coincide con el tamaño de una vía

$$\# \text{ líneas por vía} = \frac{\# \text{ líneas cache}}{\# \text{ vías}} = \frac{32 \frac{\text{líneas}}{cache}}{2 \frac{\text{vías}}{cache}} = 16 \frac{\text{líneas}}{\text{vía}} = 2^4 \frac{\text{líneas}}{\text{vía}}$$

# líneas por vía x capacidad de una línea = 
$$2^4 \frac{\text{líneas}}{\text{vía}} \times 64 \frac{B}{\text{línea}} = 2^{10} \frac{B}{\text{vía}}$$

¿Cuántos bloques entran en memoria principal?

$$\frac{\text{capacidad memoria}}{\text{tamaño bloque}} = \frac{1MB/MP}{2^{10}B/bloque} = \frac{2^{20}B/MP}{2^{10}B/bloque} = 2^{10}\frac{bloque}{MP}$$

- ¿Cuántas líneas entran en un bloque? Para esta configuración, tantas como entren en una vía: 2<sup>4</sup> (ver punto 1).
- 4 ¿Cuántas direcciones distintas refieren a una misma línea?

tamaño del tag $\times$ tamaño de la linea =  $2^{10}$  bits  $\times$   $2^{6}$  bits =  $2^{16}$  direcciones

#### ¿Cómo puedo saber si está cargada la línea donde se encuentra la palabra referida por la dirección C34A6?

Primero me fijo cuánto mide cada campo de una dirección de memoria para esta configuración de cache: **tag** 10 *bits*, **conjunto** 4 *bits*, **índice** 6 *bits*. Después, paso la dirección a binario para saber el valor de los campos correspondientes a esa dirección.

La dirección en binario:

$$\frac{C}{1100} \frac{3}{0011} \frac{4}{0100} \frac{A}{1010} \frac{6}{0110}$$

Agrupada según los campos tag, conjunto e índice:

$$\frac{30D}{11,0000,1101}$$
  $\frac{2}{0010}$   $\frac{26}{10,0110}$ 

Finalmente, me tengo que fijar si en el lugar reservado para las líneas número 0x2, está cargada la línea correspondiente al bloque número 0x30D, en cualquiera de los dos espacios del conjunto; si es así, la línea correspondiente a la dirección pedida está cargada en cache.

Memoria Principal 1 MB, direccionable a byte.

Longitud de Palabra: 2 bytes.

Cache 32 líneas de 64 bytes cada una, 2 vías, FIFO.

Dar hit-rate de las siguientes lecturas a memoria, indicando en cada paso cómo queda la cache:

0xC34A6

0xC38AB

0xC3480

0×C34D4

0xC34FF

0xC34BF

0×00090

Memoria Principal 1 MB, direccionable a byte. Longitud de Palabra: 2 bytes.

Cache 32 líneas de 64 bytes cada una, 2 vías, FIFO.

Dirección	Tag	Set	Índice	Resultado	Estado Cache	Notas
C34A6 30D 2 26 1100 0011 01 00 10 10 0110	30D	2	26	Miss	{2:30D}	cargué 2:30D
C38AB 30E 2 2B 1100 0011 10 00 10 10 1011	30E	2	2B	Miss	{2:30D, 2:30E}	cargué 2:30E
C3480 30D 2 00 1100 0011 01 00 10 00 0000	30D	2	00	Hit	{2:30D, 2:30E}	
C34D4 30D 3 14 1100 0011 01 00 11 01 0100	30D	3	14	Miss	{2:30D, 2:30E} {3:30D}	cargué 3:30D
C34FF 30D 3 3F 1100 0011 01 00 11 11111	30D	3	3F	Miss	{2:30D, 2:30E} {3:30D} {4:30D}	acc. desalineado; cargué 4:30D
C34BF 30D 2 3F 1100 0011 01 00 10 11 1111	30D	2	3F	Hit	{2:30D, 2:30E} {3:30D} {4:30D}	acc. desalineado; tengo ambas
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	000	2	10	Miss	{2:30E, 2:000} {3:30D} {4:30D}	desalojé 2:30D

Hit rate =  $\frac{2}{7} \approx 0.29 \%$ 

## Ejercicio 4 - Tipo parcial

Se estudia agregar una memoria cache a una computadora cuyas palabras y direcciones de memoria son de 16 *bits*, y que trabaja con *direccionamiento* a byte.

Hasta el momento se barajan dos opciones: una cache asociativa por conjuntos de 4 vías, y una de correspondencia directa (o *mapeo* directo).

Cada cache puede almacenar hasta **1024** bytes de información (sin contar el espacio necesario para los *tags*) y se organiza en 64 líneas en **total**.

Indique cómo se distribuyen los bits de una dirección de memoria en los campos correspondientes para cada una de las caches mencionadas.

## Ejercicio 4 - Tipo parcial

Se conoce que este fragmento de código insume gran parte del tiempo de cómputo. Junto a cada instrucción se indica el acceso a memoria necesario para el fetch de la instrucción, no así los necesarios para acceder a los datos.

	Código	Pedido a memoria del <i>fetch</i>		
	MOV R6, 0×10	0x9C13		
	MOV R2, 0x801A	0x9C15, 0x9C17		
mejillón:	ADD R5, [R2]	0x9C19		
	ADD R5, $[R2 + 0 \times 15]$	0x9C1B		
	SUB R6, 0×01	0x9C1D		
	CMP R6, 0×00	0×9C1F		
	JNE mejillón	0x9C21		

Simule los accesos a memoria que realiza este programa hasta la comparación (inclusive) utilizando la cache de **mapeo directo**. **Tenga en cuenta los accesos a datos, que no fueron detallados en la tabla**. Indique en cada paso el contenido de la cache, y cuando corresponda detalle si se producen *hits, misses*, desalojos (señalando la línea desalojada) y/o accesos desalineados. ¿Cuál es el *hit rate* de esta ejecución parcial?

¿Cuál de las dos caches es más conveniente para la ejecución completa del fragmento del programa presentado?

#### Ejercicio 4 - Tipo parcial (a) Correspondencia directa

- ¿Cuánto mide un bloque para esta configuración? Para esta configuración, coincide con el tamaño de la cache, es decir 2<sup>10</sup> B (lo dice el enunciado)
- ¿Cuántos bloques entran en memoria principal?

$$\frac{\text{capacidad memoria}}{\text{tamaño bloque}} = \frac{\# \text{ direcciones} \times \text{tamaño unidad dir.}}{\text{tamaño bloque}} = \frac{2^{16} \times 1B}{2^{10}B} = 2^6$$

¿Cuántas líneas entran en un bloque? Para esta configuración, tantas como entren en la cache:

capacidad de una línea = 
$$\frac{\text{capacidad cache}}{\# \text{ líneas en cache}} = \frac{2^{10} B}{64} = \frac{2^{10} B}{2^6} = 2^4 B$$

$$\frac{\text{capacidad cache}}{\text{capacidad de una línea}} = \frac{2^{10}B/\text{cache}}{2^{4}B/\text{linea}} = 2^{6}\frac{\text{linea}}{\text{cache}}$$

¿Cuántas direcciones distintas refieren a una misma línea?

tamaño del tag x tamaño de la linea =  $2^6$  bits x  $2^4$  bits =  $2^{10}$  direcciones

Por lo tanto, la decodificación de la dirección de 16 bits será

tag		línea	índice	
	6 bits	6 bits	4 bits	

#### Ejercicio 4 - Tipo parcial (a) Asociativa por conjuntos

¿Cuánto mide un bloque para esta configuración? Para esta configuración, coincide con el tamaño de una vía.

$$\frac{\text{capacidad cache}}{\# \text{ vías}} = \frac{2^{10}B}{4} = 2^8B$$

¿Cuántos bloques entran en memoria principal?

$$\frac{\text{capacidad memoria}}{\text{tamaño bloque}} = \frac{\# \text{ direcciones} \times \text{tamaño unidad dir.}}{\text{tamaño bloque}} = \frac{2^{16} \times 1B}{2^8 B} = 2^8$$

- **¿Cuántas líneas entran en un bloque?** Hay 64 líneas, que se dividen entre las 4 vías. Por lo tanto, hay  $\frac{64}{4} = 16 = 2^4$  conjuntos.
- 4 ¿Cuántas direcciones distintas refieren a una misma línea?

tamaño del tag $\times$ tamaño de la linea =  $2^8$  bits  $\times$   $2^4$  bits =  $2^{12}$  direcciones

Por lo tanto, la codificación será

tag	conjunto	índice		
8 bits	4 bits	4 bits		

### Ejercicio 4 - Tipo parcial (b) Accesos a memoria - Acceso directo

Dirección	Tag	Línea	Índice	Resultado	Estado Cache	Notas
0×9C13 27 01 3 10 0111 00 0001 0011	27	1	3	Miss	{1: 27}	cargué 1:27
0×9C15 27 01 5 10 0111 00 0001 0101	27	1	5	Hit	{1: 27}	
0x9C17 27 01 7 10 0111 00 0001 0111	27	1	7	Hit	{1: 27}	
0×9C19 27 01 9 10 0111 00 0001 1001	27	1	9	Hit	{1: 27}	
0×801A 20 01 A 10 0000 00 0001 1010	20	1	A	Miss	{1: 20}	desalojé 1: 27; cargué 1:20
0x9C1B 27 01 B 10 0111 00 0001 1011	27	1	В	Miss	{1: 27}	desalojé 1: 20; cargué 1:27
0x802F 20 02 F 10 0000 00 0010 1111	20	2	F	Miss	{1: 27} {2: 20} {3: 20}	acc. desalineado; cargué 2: 20; cargué 3: 20
0×9C1D 27 01 D 10 0111 00 0001 1101	27	1	D	Hit		
0×9C1F 27 01 F 10 0111 00 0001 F 1111	27	1	F	Miss	{1: 27} {2: 27} {3: 20}	acc. desalineado; desalojé 2:20; cargué 2: 27

Hit rate =  $\frac{4}{9} \approx 0.44 \%$ 

#### Ejercicio 4 - Tipo parcial (c) Comparación

Para este bloque de código sería más conveniente la caché asociativa por conjuntos de 4 vías, ya que se eliminarían los desalojos gracias al más alto nivel de asociatividad.

# ¿Preguntas?



# ¿Preguntas?



Max Headroom