Escuela de
Ingeniería y Arquitectura
Universidad Zaragoza

Departamento de
Informática e Ingeniería
de Sistemas
Universidad Zaragoza

Tema 6 – Organización de una cache: Correspondencia

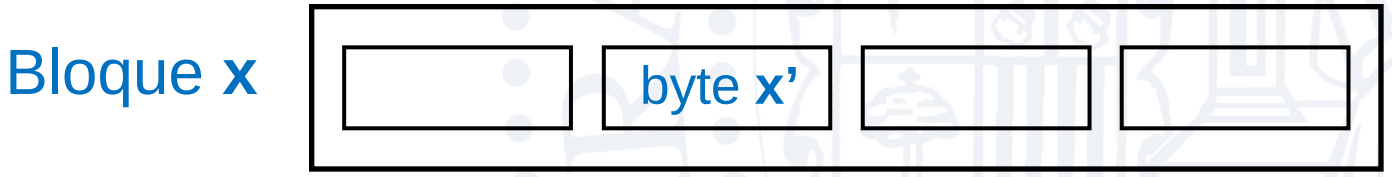
P. Ibáñez, J.L. Briz, V. Viñals, J. Alastruey, J. Resano
Arquitectura y Tecnología de Computadores
Departamento de Informática e Ingeniería de Sistemas

Guión del tema

- Organización de la memoria en bloques
- Descripción de alternativas de correspondencia
- Implementación
- Modelo de las 3 Cs

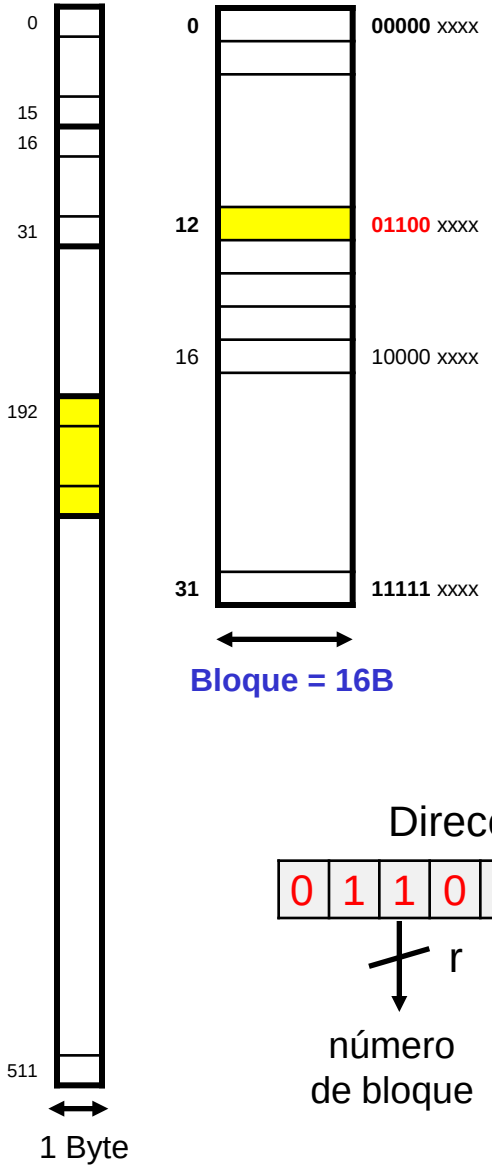
Organización de la memoria en bloques

- Bloques alineados
- Unidad de correspondencia entre Mp y Mc
  - nº de bloque: clave para buscar un bloque en Mc
- Unidad de transferencia entre Mp y Mc
  - Bloques enteros fluyen desde Mp hacia Mc en caso de fallo de cache



Correspondencia totalmente asociativa

Memoria Principal: 512 Bytes



Cache 128B

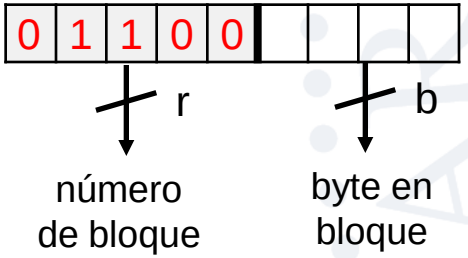
Bloque	Tag	V	U
	01100	1	

5 bits

- Cualquier bloque puede estar en cualquier contenedor
- La búsqueda se realiza en paralelo
  - un comparador por contenedor
- Necesario un algoritmo de reemplazo

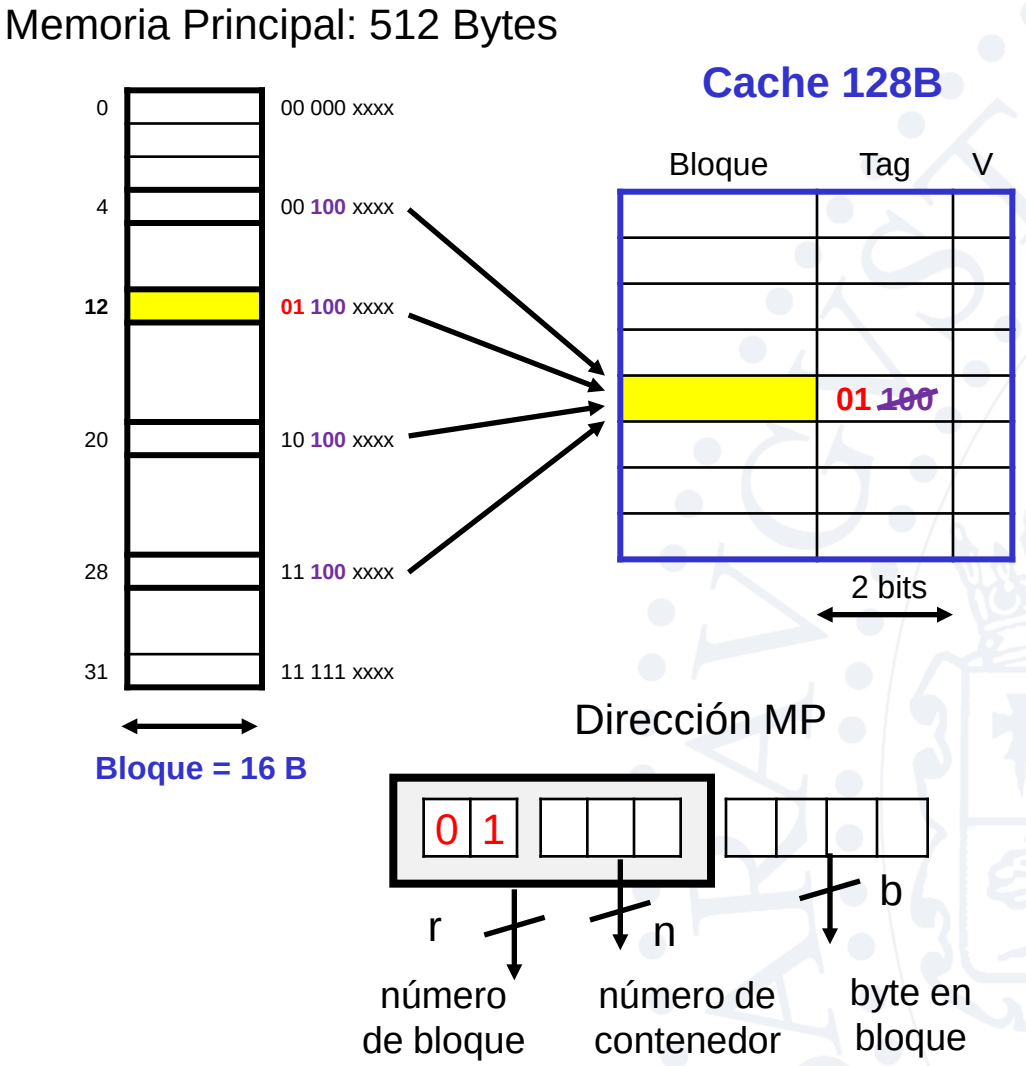
- V: bit de validez
- U: bits de uso.
  → bloque víctima (a reemplazar)

Dirección MP



- ¿Implementación?
  - Transf. con el procesador: 4B
  - Transferencia con memoria: 4-16B
  - SRAM contenidos:
    - 8 x 16 B o 32 x 4 B
  - ¿Reemplazo?

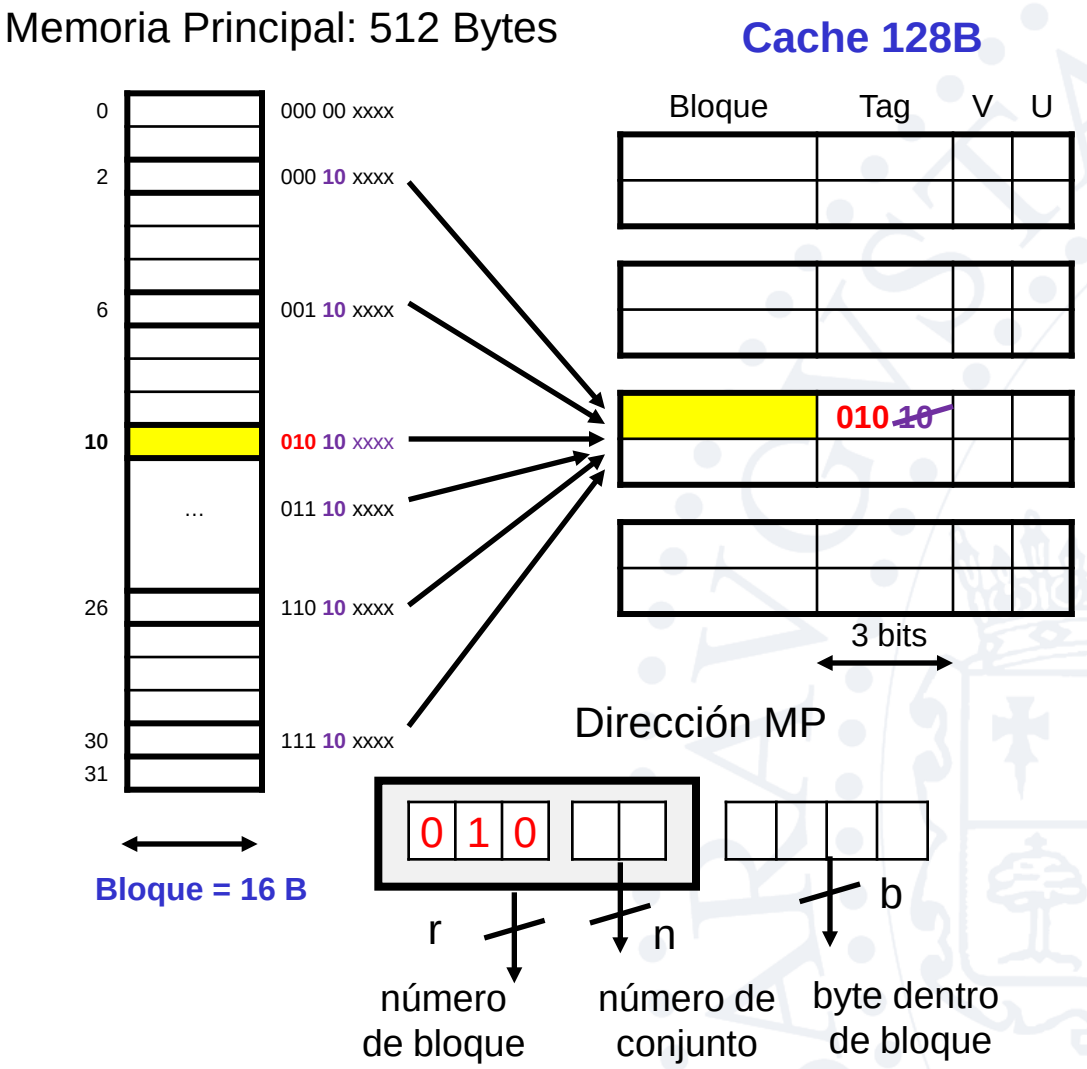
Correspondencia directa



- Cada bloque de Mp sólo puede estar en **un contenedor** o **conjunto** de Mc
- Conflictos: aumentan la tasa de fallos
- Sin algoritmo de reemplazo
- Implementación
  - Se leen en paralelo tag y bloque
  - Un sólo comparador para toda la cache
  - A igualdad de tamaño, correspondencia con Tag menores

¿Implementación?

Correspondencia asociativa por conjuntos



- **N** conjuntos de tamaño **As contenedores** (o **vías**)
- Cada bloque de Mp sólo puede estar en un **conjunto** de Mc
- En un conjunto: colocación libre
- Solución de compromiso

128B = N x As x L
4 conjuntos
x 2 bloques/conjunto
x 16 Bytes/bloque

¿Implementación?

Correspondencias

- Ejercicio: calcular la capacidad del tercer nivel de cache del Intel 6-core Xeon 7400 (Dunnington) sabiendo que se compone de 4 bancos, y que cada uno tiene las siguientes características:
  - 4096 conjuntos
  - Asociatividad 16
  - Tamaño de bloque: 64 bytes
- Repite el cálculo para asociatividad 12 y 8 (correspondiente a procesadores de menor coste)

Correspondencias

- Ejercicio: suponiendo para el direccionamiento de bytes en memoria se utilizan 32 bits (4 GBytes direccionables), realizar la descomposición de una dirección en los campos necesarios para direccionar estas caches (r, n , b):
  - C = 64 KB, asoc. 8, bloque 32 bytes
  - C = 64 KB, asoc. 1, bloque 32 bytes

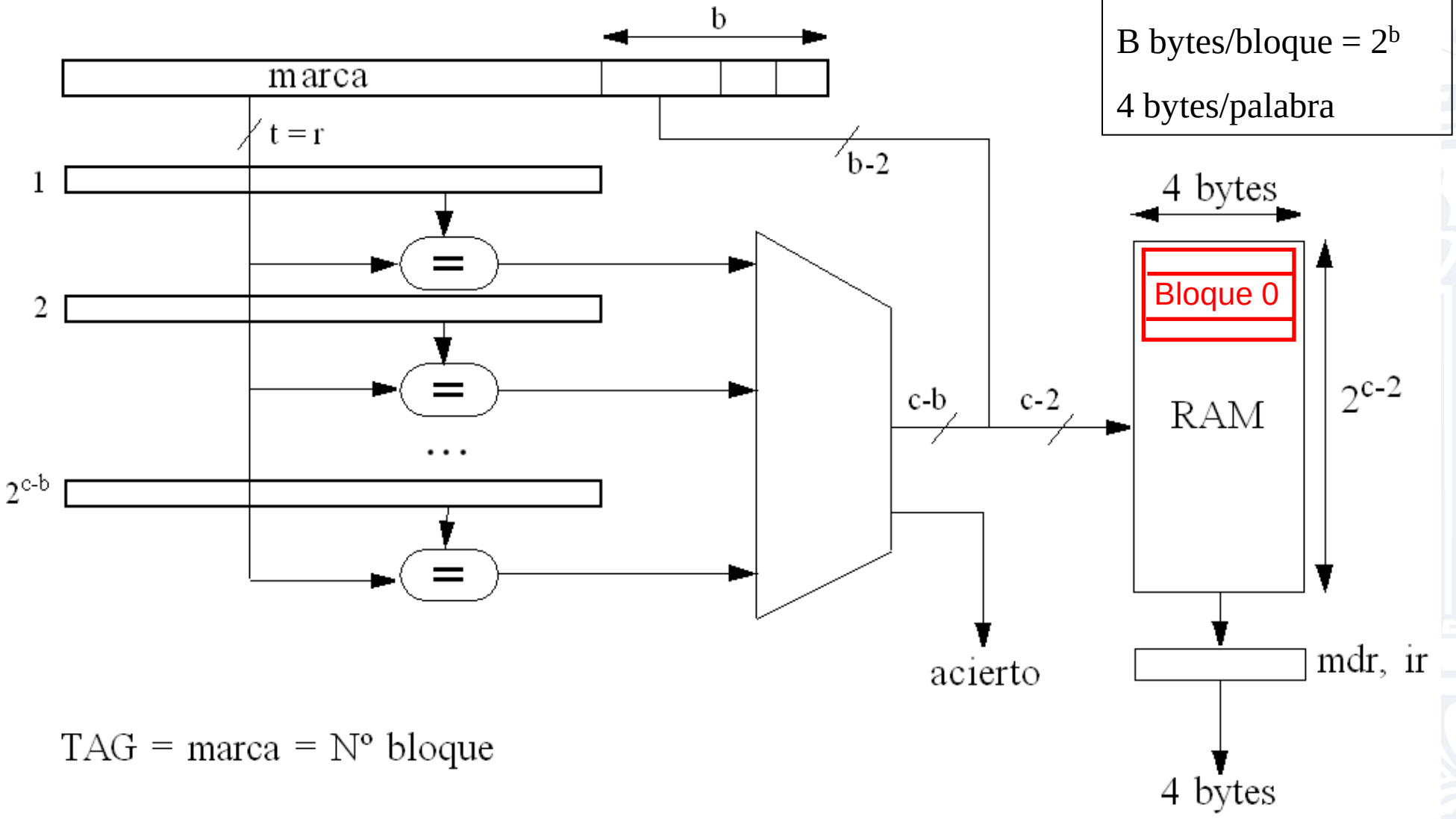
Correspondencias

- Ejercicio: un procesador ejecuta un bucle que repite la secuencia de direcciones siguiente dos veces:
  - 0xA0, 0x34, 0x74, 0xA8, 0x30

  Suponiendo que la cache está inicialmente vacía (todos los bloques inválidos), obtener la tasa de fallos para:
  - Mp = 256 bytes
  - Cache = 64 bytes
  - Bloques de 16 bytes
  - Asociatividad = {1, 2, 4}

Implementación

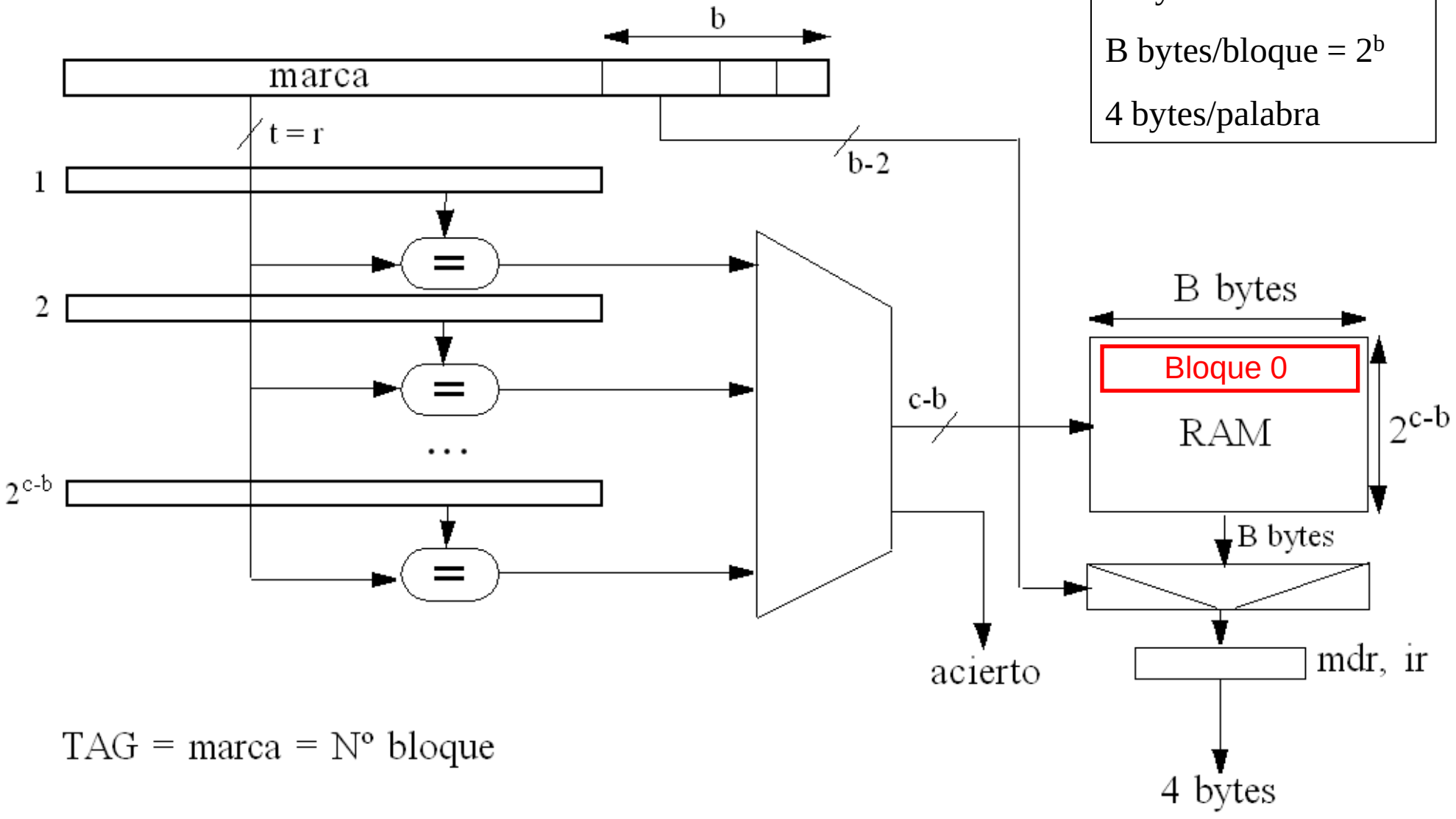
- Completamente asociativa: TTL (1)



TAG = marca = Nº bloque

Implementación

- Completamente asociativa: TTL (2)



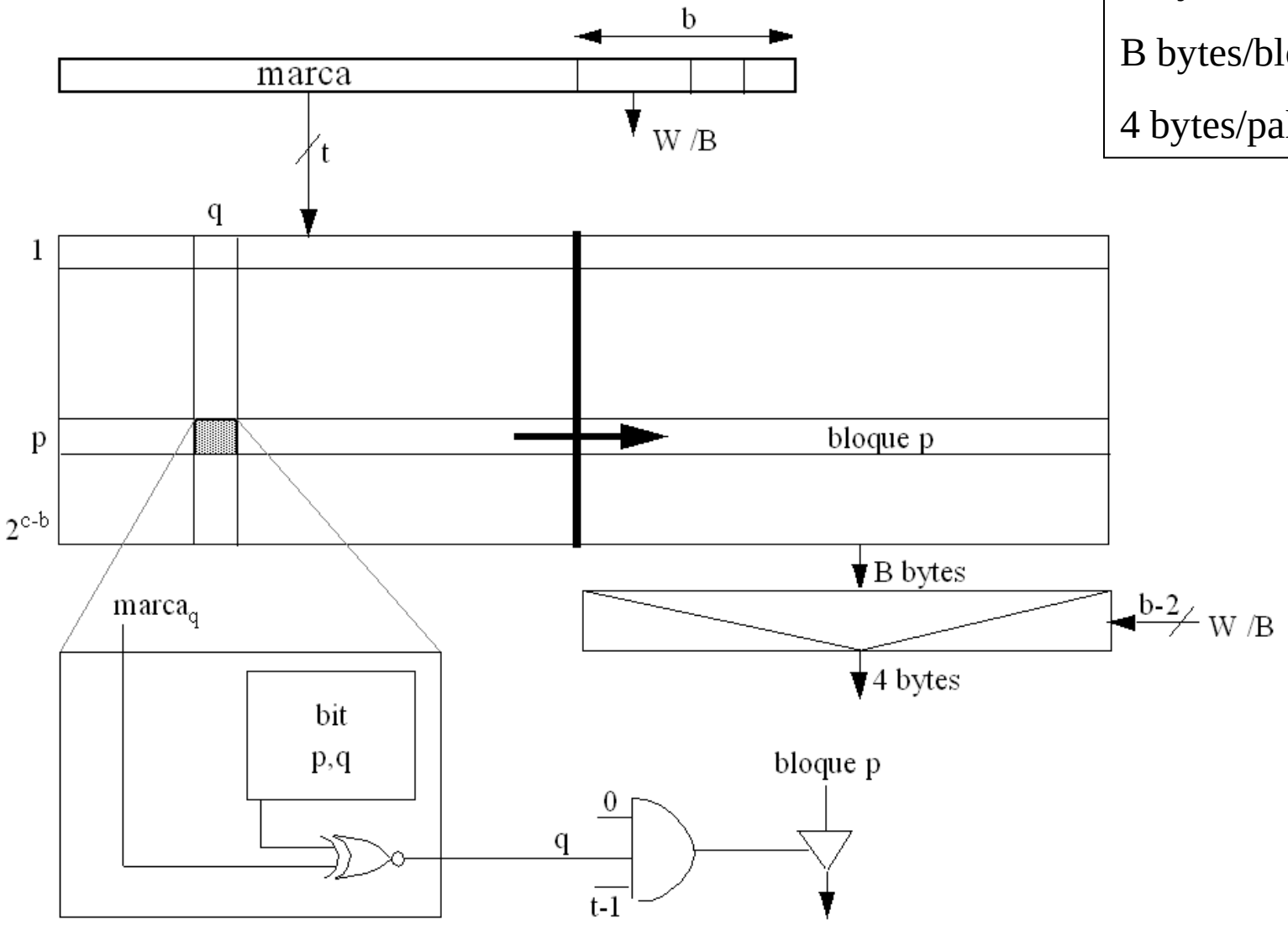
Implementación

- Completamente asociativa: CAM (VLSI)

C bytes = 2^c

B bytes/bloque = 2^b

4 bytes/palabra



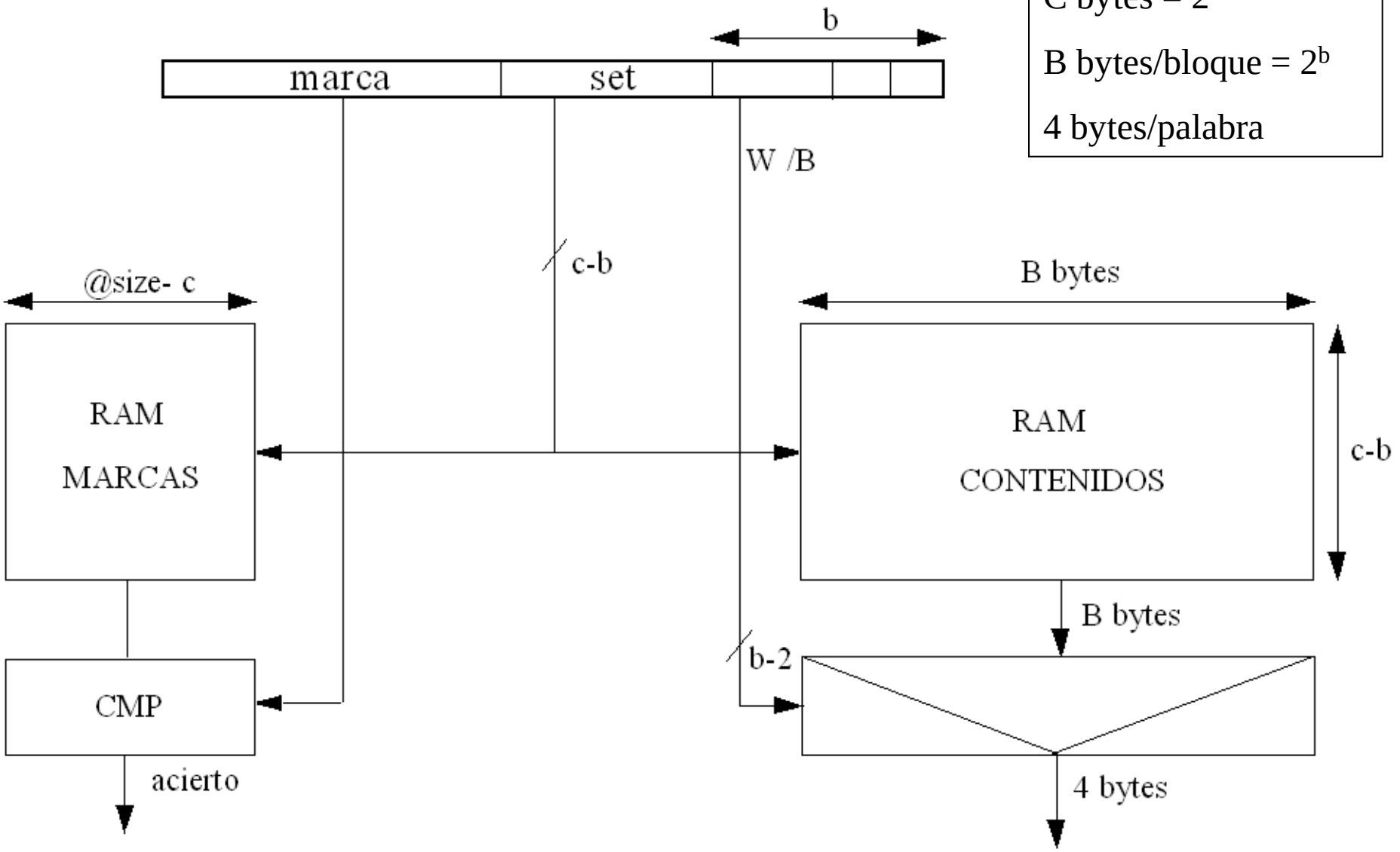
Implementación

- Correspondencia directa (acceso paralelo)

C bytes = 2^c

B bytes/bloque = 2^b

4 bytes/palabra



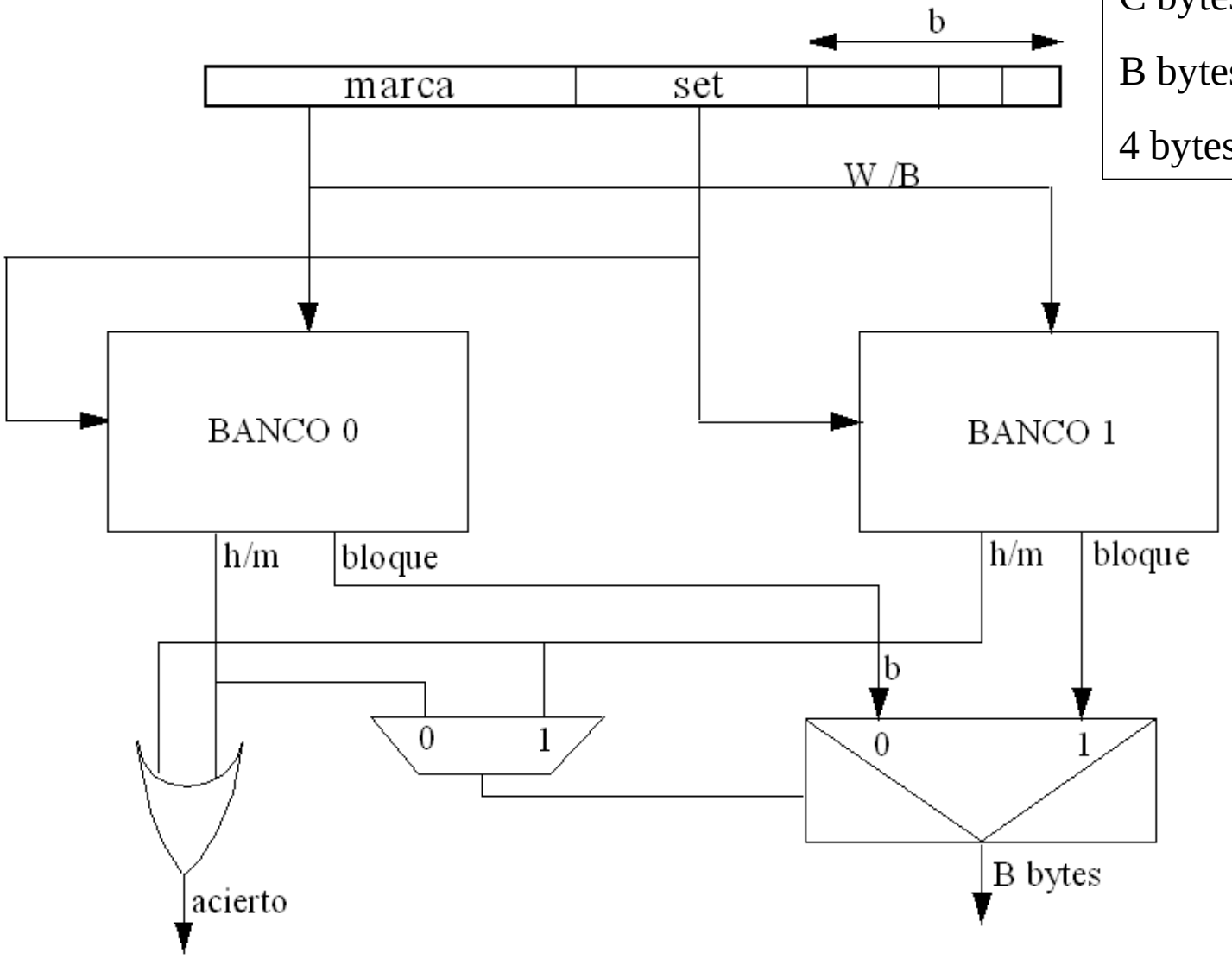
Implementación

- Asociativa por conjuntos (acceso paralelo)

C bytes = 2^c

B bytes/bloque = 2^b

4 bytes/palabra



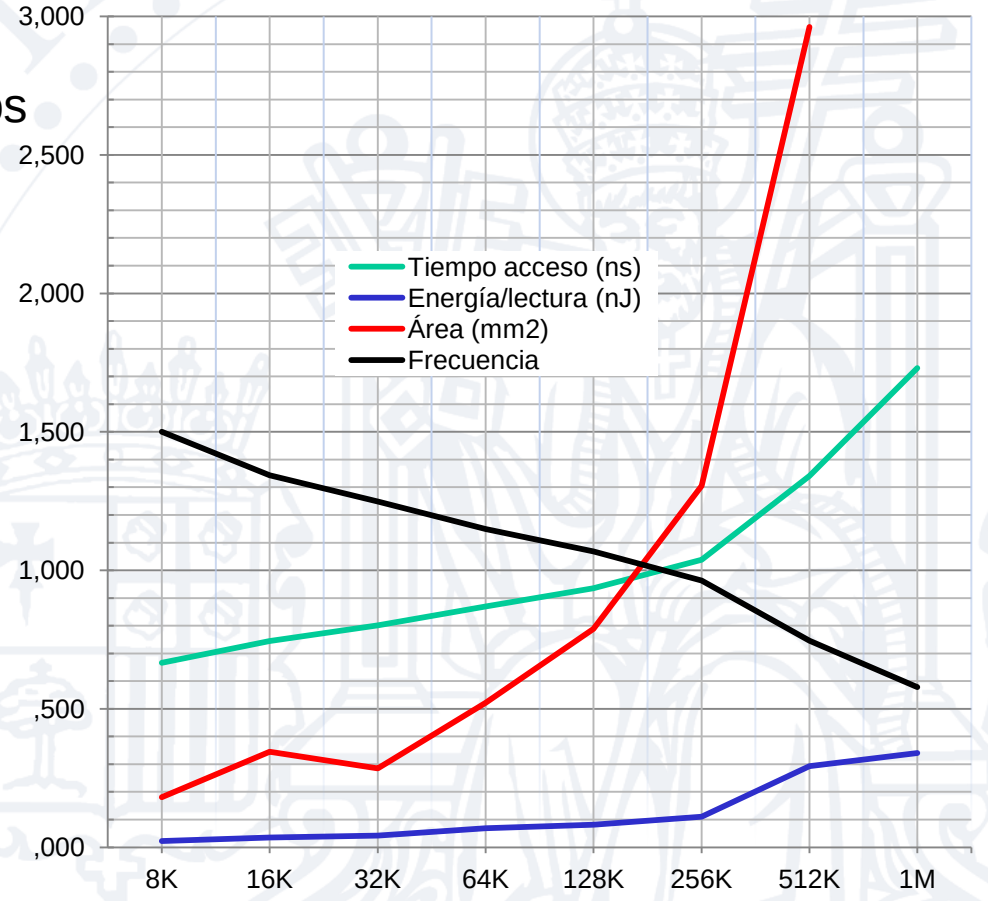
Implementación: problema

- En la cache asociativa por conjuntos de acceso paralelo se accede en paralelo a Marcas y Contenidos en todas las vías a la vez, lo cual consume mucha energía, pero a cambio es la opción más rápida
- Para asociatividades elevadas y caches de nivel 2 ó 3, la latencia es menos crítica, pero el consumo es más
- Proponed una implementación alternativa, llamada de acceso secuencial, que sea más lenta, pero que consuma menos

Implementación

- Análisis cache: retardo, área y consumo
  - V1.0 http://www.ece.ubc.ca/~stevew/cacti/run_frame.html
  - V5.2 http://quid.hpl.hp.com:9082/cacti/index.y

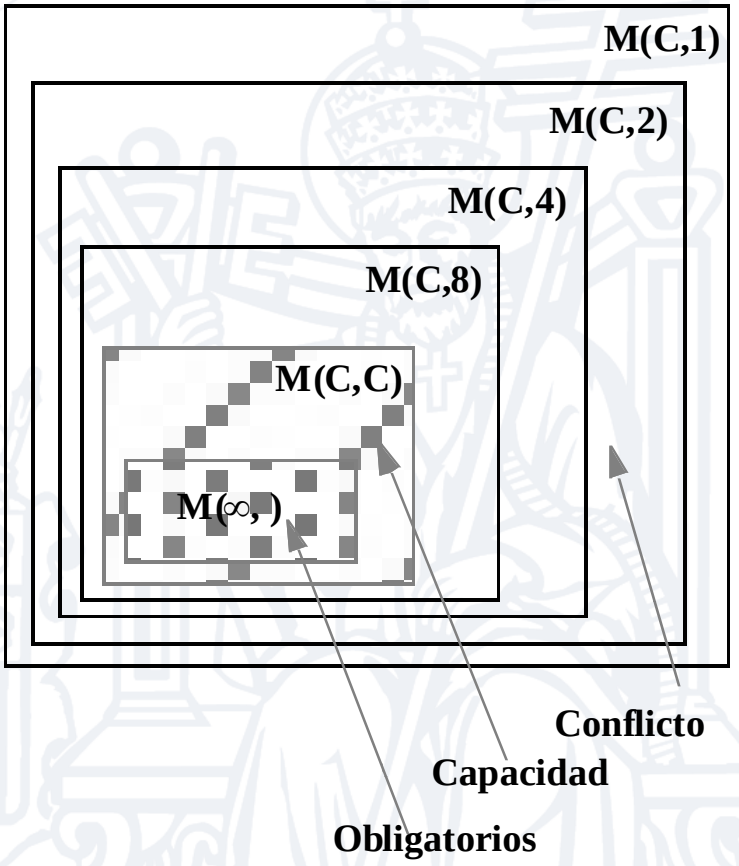
  escala bien transistores y cables
  redes en H para interconectar bancos
  soporta varios escenarios:
  - low power, high performance, …
  modela DRAMs y SRAMs



Modelo de las 3 Cs: tipos de fallos

- Fallos obligatorios (*Compulsory*): M(∞,)
  - Primer acceso a un bloque
  - Fallos de una cache infinita
- Fallos de capacidad
  - Tamaño programa > cache
    - código y/o datos
  - Aparecen al limitar el tamaño, pero en una cache completamente asociativa
- Fallos de conflicto
  - Los bloques compiten por conjuntos de asociatividad limitada

M (C, As)=
nº fallos (Capacidad, Asociatividad)

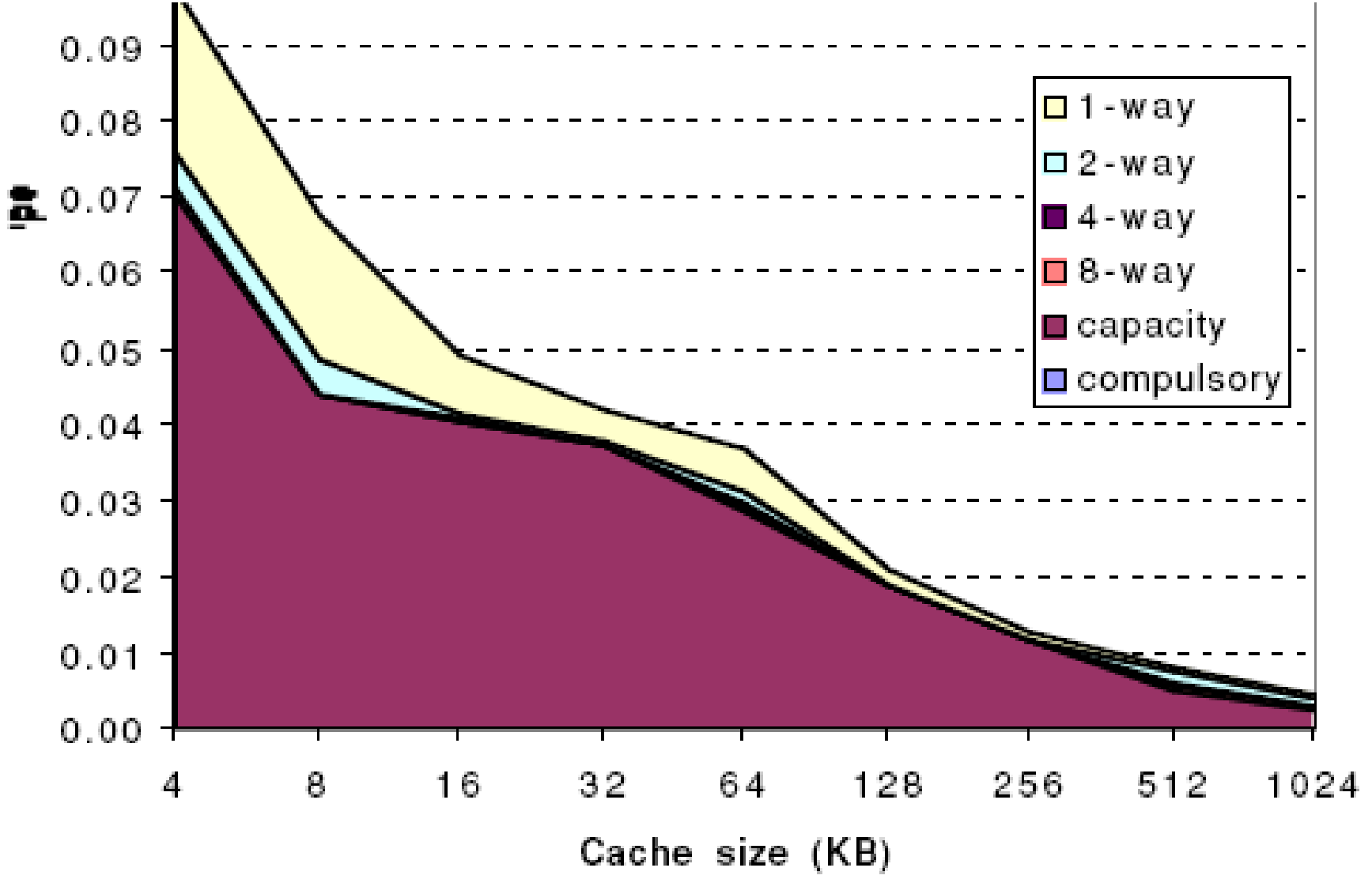


Cache size (KB)	Degree associative	Total miss rate	Miss rate components (relative percent) (*sum = 100% of total miss rate*)					
			Compulsory		Capacity		Conflict	
4	1-way	0.098	0.0001	0.1%	0.070	72%	0.027	28%
4	2-way	0.076	0.0001	0.1%	0.070	93%	0.005	7%
4	4-way	0.071	0.0001	0.1%	0.070	99%	0.001	1%
4	8-way	0.071	0.0001	0.1%	0.070	100%	0.000	0%
8	1-way	0.068	0.0001	0.1%	0.044	65%	0.024	35%
8	2-way	0.049	0.0001	0.1%	0.044	90%	0.005	10%
8	4-way	0.044	0.0001	0.1%	0.044	99%	0.000	1%
8	8-way	0.044	0.0001	0.1%	0.044	100%	0.000	0%
16	1-way	0.049	0.0001	0.1%	0.040	82%	0.009	17%
16	2-way	0.041	0.0001	0.2%	0.040	98%	0.001	2%
16	4-way	0.041	0.0001	0.2%	0.040	99%	0.000	0%
16	8-way	0.041	0.0001	0.2%	0.040	100%	0.000	0%
32	1-way	0.042	0.0001	0.2%	0.037	89%	0.005	11%
32	2-way	0.038	0.0001	0.2%	0.037	99%	0.000	0%
32	4-way	0.037	0.0001	0.2%	0.037	100%	0.000	0%
32	8-way	0.037	0.0001	0.2%	0.037	100%	0.000	0%
64	1-way	0.037	0.0001	0.2%	0.028	77%	0.008	23%
64	2-way	0.031	0.0001	0.2%	0.028	91%	0.003	9%
64	4-way	0.030	0.0001	0.2%	0.028	95%	0.001	4%
64	8-way	0.029	0.0001	0.2%	0.028	97%	0.001	2%
128	1-way	0.021	0.0001	0.3%	0.019	91%	0.002	8%
128	2-way	0.019	0.0001	0.3%	0.019	100%	0.000	0%
128	4-way	0.019	0.0001	0.3%	0.019	100%	0.000	1%
128	8-way	0.019	0.0001	0.3%	0.019	100%	0.000	0%
256	1-way	0.013	0.0001	0.5%	0.012	94%	0.001	6%
256	2-way	0.012	0.0001	0.5%	0.012	99%	0.000	0%
256	4-way	0.012	0.0001	0.5%	0.012	99%	0.000	0%
256	8-way	0.012	0.0001	0.5%	0.012	99%	0.000	0%
512	1-way	0.008	0.0001	0.8%	0.005	66%	0.003	33%
512	2-way	0.007	0.0001	0.9%	0.005	71%	0.002	28%
512	4-way	0.006	0.0001	1.1%	0.005	91%	0.000	8%
512	8-way	0.006	0.0001	1.1%	0.005	95%	0.000	4%

HePa2003 pág.424.- Tasa total de fallos para cada tamaño de Mc y porcentaje de cada uno según el modelo de las tres C's. Los obligatorios son independientes del tamaño de Mc. Los de capacidad decrecen cuando el tamaño de Mc aumenta. Los de conflicto bajan al aumentar la asociatividad. Observar que se cumple la regla del 2:1 hasta tamaños de 128 KB: una Mc de mapeo directo de tamaño N tiene aprox. la misma tasa de fallos que una 2-way de tamaño N/2. Datos de SPECint y SPECfp; reemplazo LRU.

Modelo de las 3 Cs

- Un ejemplo experimental



Modelo de las 3 Cs

- Obtención