Características compartidas por la jerarquía de memoria

Característica	Valores para L1	Valores para L2
Tamaño total en bloques	250-2000	2500-25000
Tamaño total en kilobytes	16-64	125-2000
Tamaño del bloque en bytes	16-64	64-128
Miss penalty en ciclos	10-25	100-1000
Miss rate	2%-5%	0,1%-2%

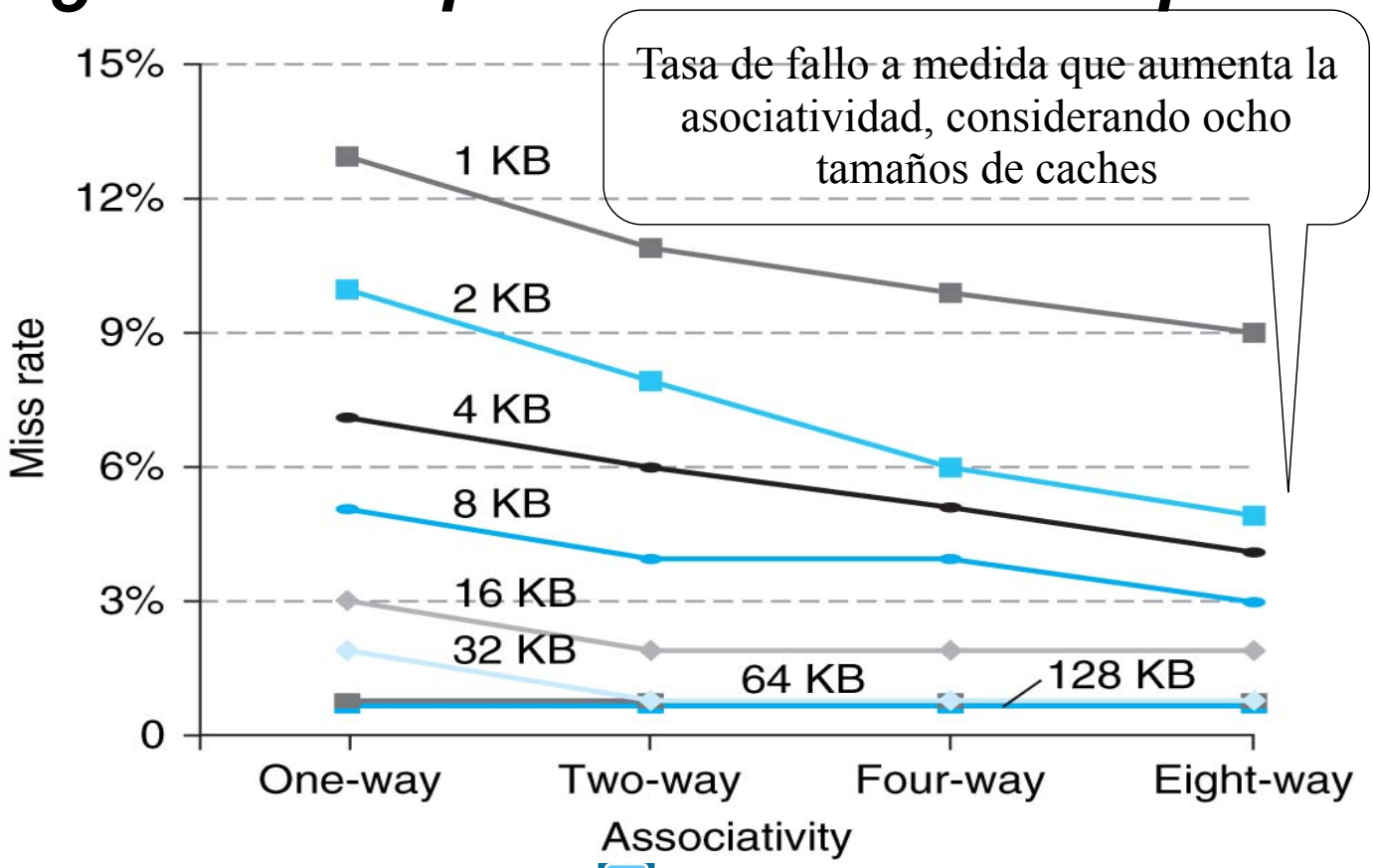
- ¿Dónde se puede ubicar un bloque?
- ¿Cómo se encuentra un bloque?
- ¿Qué bloque reemplazar en un miss?
- ¿Qué ocurre con las escrituras?
- Las tres C´s: modelo de comportamiento de la jerarquía.

¿Dónde se puede ubicar un bloque?

Variantes sobre el esquema de conjunto asociativo

Nombre del esquema	Número de conjuntos	Bloques por conjunto
Correspondencia directa	Nro de bloque en cache	1
Conjunto asociativo	$\frac{\text{Nro de bloque en cache}}{\text{Asociatividad}}$	Asociatividad(2-16)
Totalmente asociativa	1	Nro de bloque en cache

¿Dónde se puede ubicar un bloque?



¿Cómo encontrar un bloque?

Ubicación de los bloques

Asociatividad	Método de ubicación	Número de comparaciones
Correspondencia directa	Índice	1
Conjunto asociativo	Indexar el conjunto y buscar entre los elementos	Asociatividad(2-16)
Completamente Asociativo	Buscar en todas las entradas	Nro de bloques en cache

- El esquema usado depende del costo del miss y del hard.
- Usar L2 permite mayor asociatividad.
- Totalmente Asociativas para caches pequeñas, con:
  - pocos comparadores
  - mejoras significativas

¿Qué bloque reemplazar?

- Principales estrategias: Aleatorio y LRU
- LRU (aproximado) en cuatro vías se pueden usar 2 bits:
  - 1 para un par de bloques LRU.
  - 1 para el bloque LRU en el par.
- Con caches de mayor asociatividad:
  - Algoritmo simple en hardware. Miss rate: Random>LRU (1,1)
  - En las grandes ambas fallan

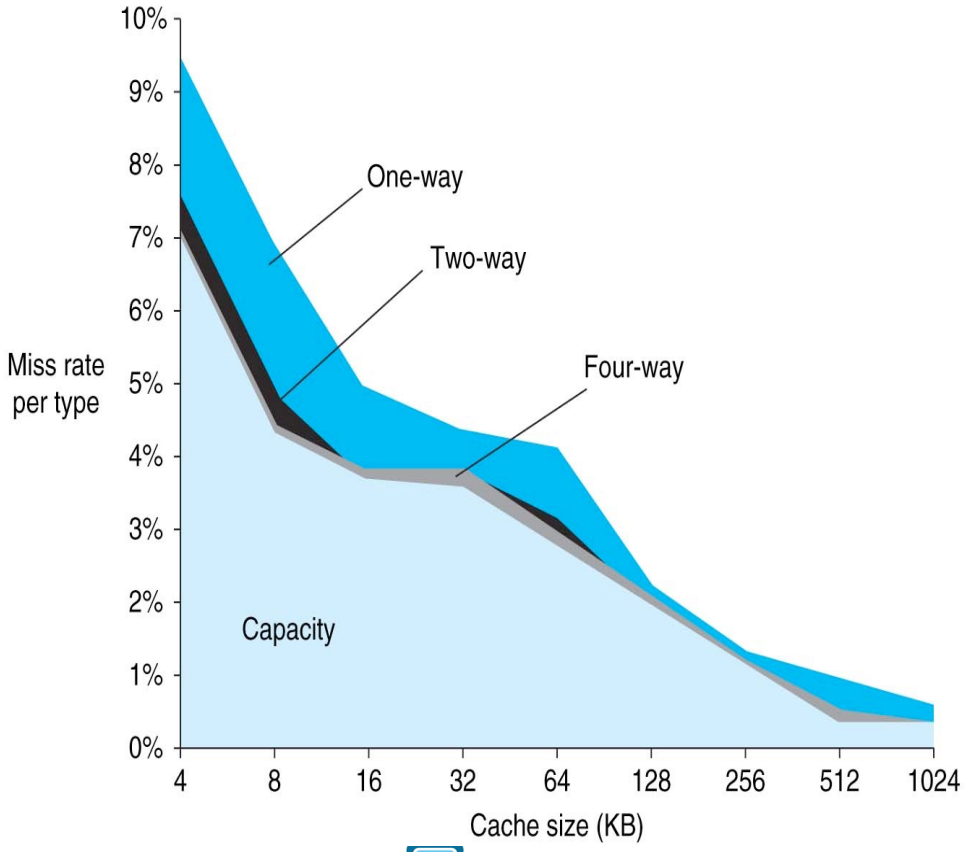
¿Qué ocurre en una escritura?

- Ventajas de write-back
  - Las palabras se escriben a la tasa de la cache.
  - Múltiples escrituras requieren un sólo acceso a memoria.
  - Puede escribir el bloque completo (high-bandwidth).
- Ventajas de write-through
  - Misses simples y baratos. Nunca escribe un bloque entero.
  - Más fácil de implementar. Usará un buffer de escritura.
- Memoria virtual sólo usa write-back porque:
  - Gran latencia al escribir en el nivel inferior.
  - La memoria no soporta la tasa de escritura del procesador.

Las tres C´s

- Origen de los misses.
- Cómo los cambios afectan los misses.
- Compulsory (Obligatorios) misses.
  - Primer acceso. Comienzo en frío
- Capacity (Por capacidad) misses.
  - No hay espacio suficiente. Bloques reemplazados y recargados.
- Conflict (Por conflictos) misses.
  - Conflictos por conjuntos de bloques. No se darían en Full Asociat.

Las tres fuentes de misses



Desafíos del diseño de la jerarquía

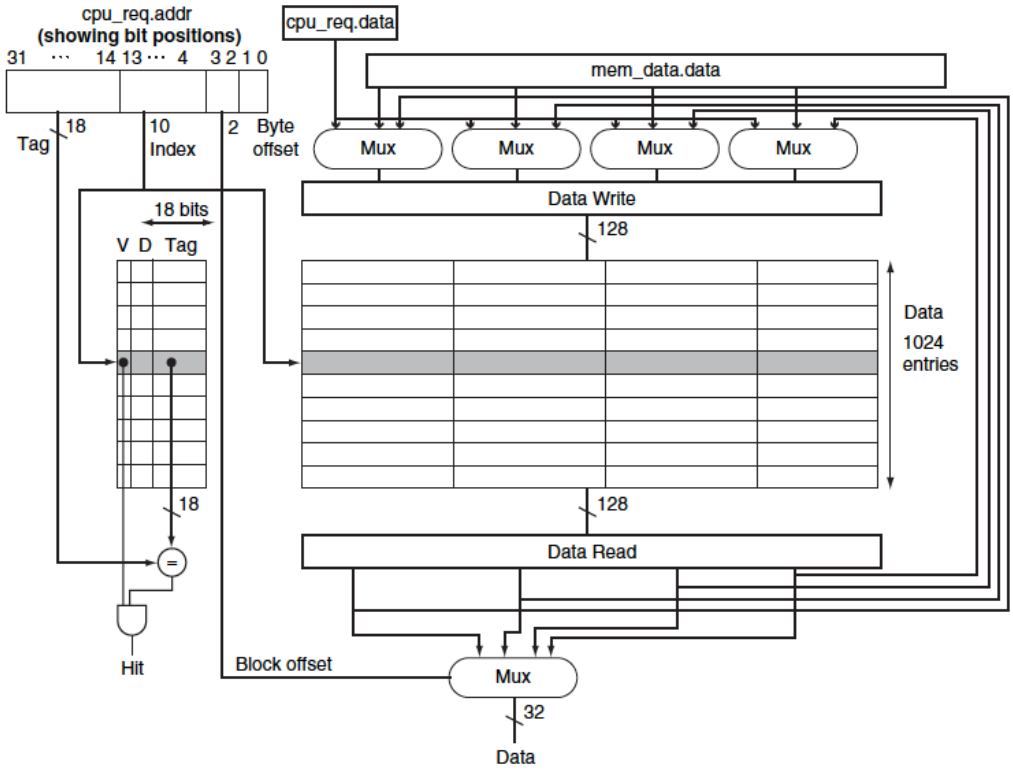
Design change	Effect on miss rate	Possible negative performance effect
Increase cache size	Decreases capacity misses	May increase access time
Increase associativity	Decreases miss rate due to conflict misses	May increase access time
Increase block size	Decreases miss rate for a wide range of block sizes due to spatial locality	Increases miss penalty. Very large block could increase miss rate

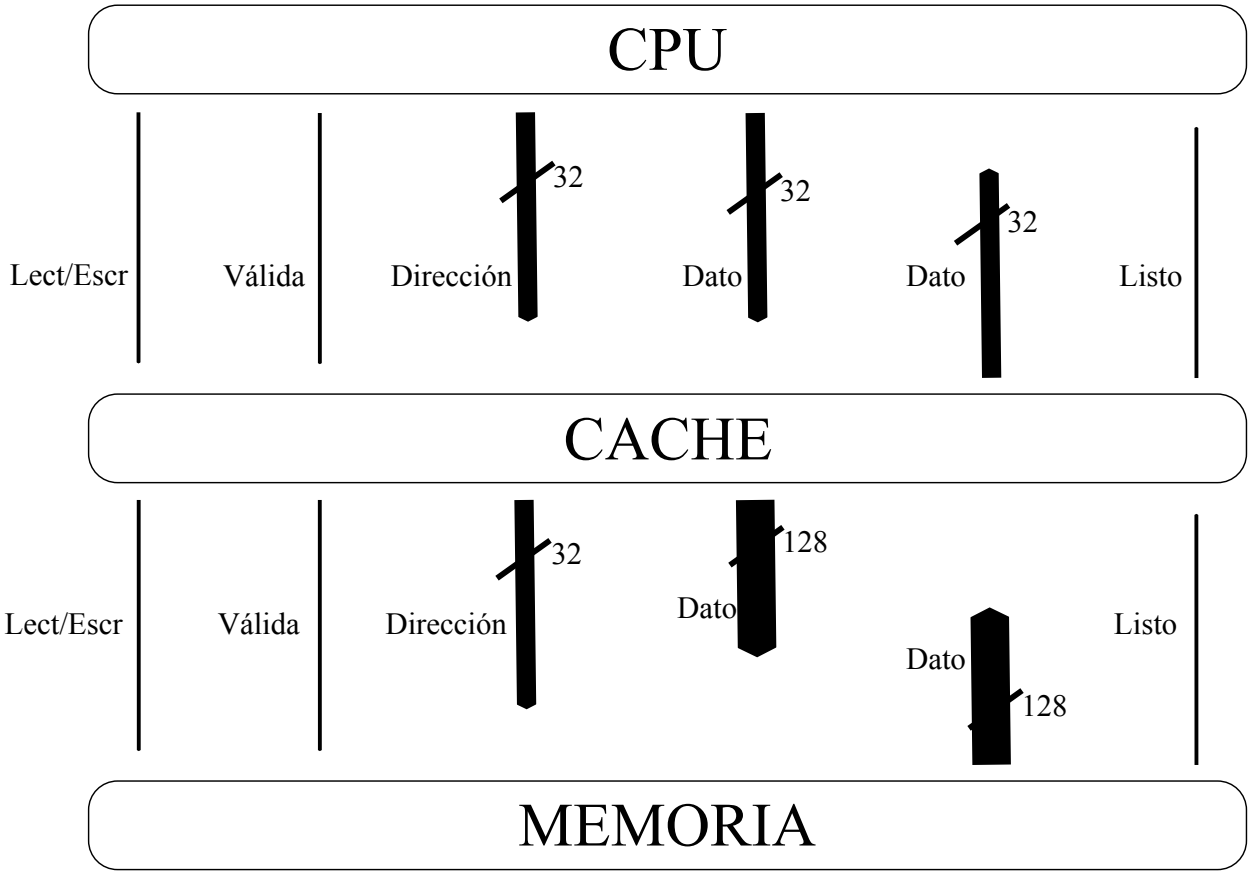
- Los cambios para mejorar la tasa de fallos pueden afectar todo el desempeño.
- La combinación de efectos positivos y negativos hace interesante el diseño de la jerarquía de memoria.

Diseño del controlador para una Cache simple

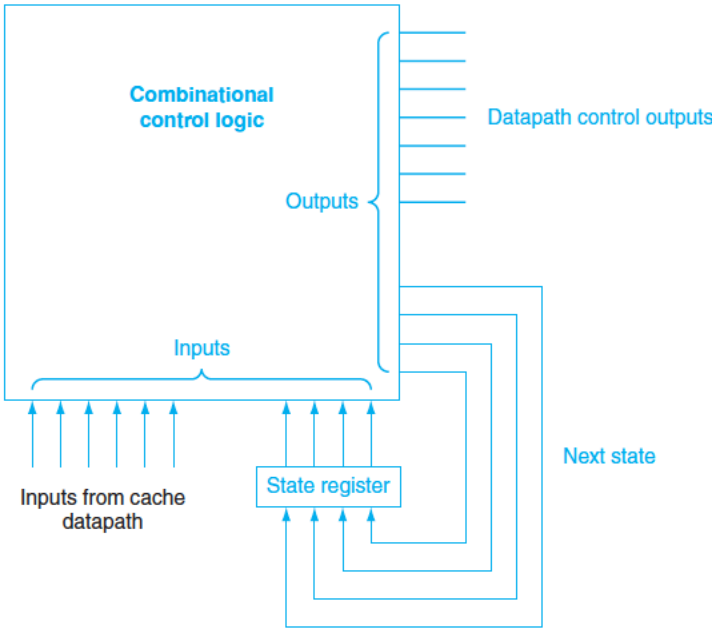
- Correspondencia Directa.
- Escritura demorada usando alojamiento del bloque.
- Tamaño del bloque 4 palabras.
- Tamaño de cache 16KiB, mantiene 1024 bloques.
- Direcciones de 32 bits.
- Cada bloque tiene bit de validez y modificado.

Diagrama de una cache simple

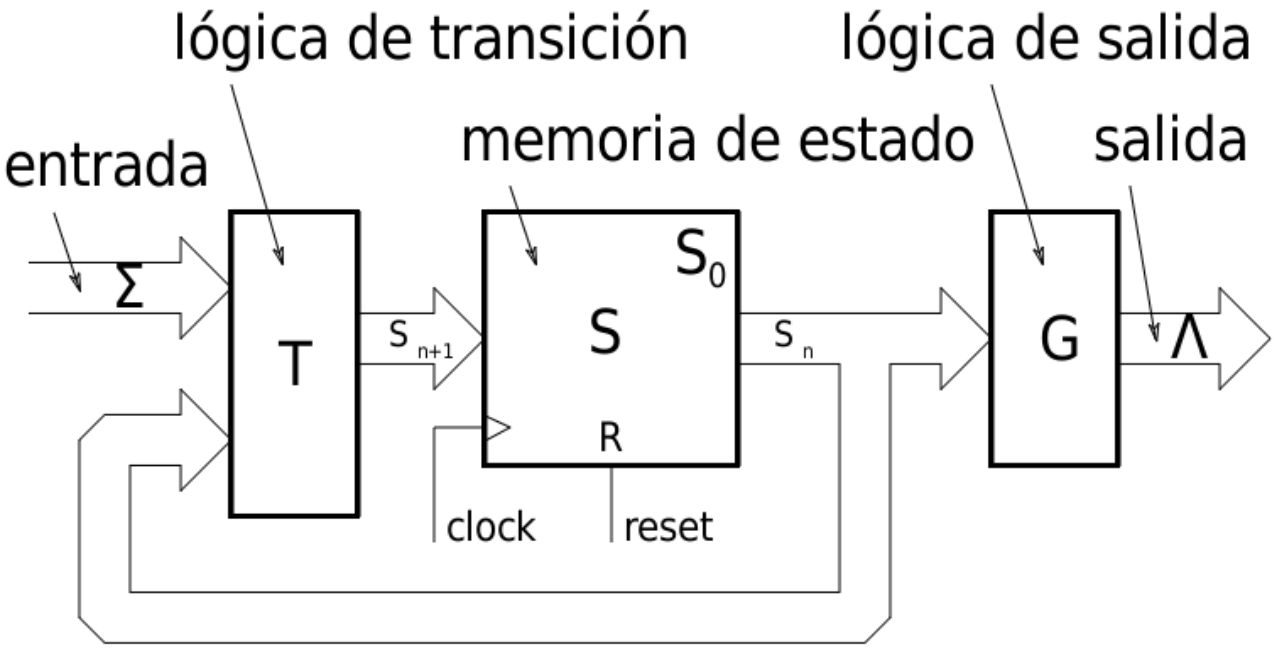




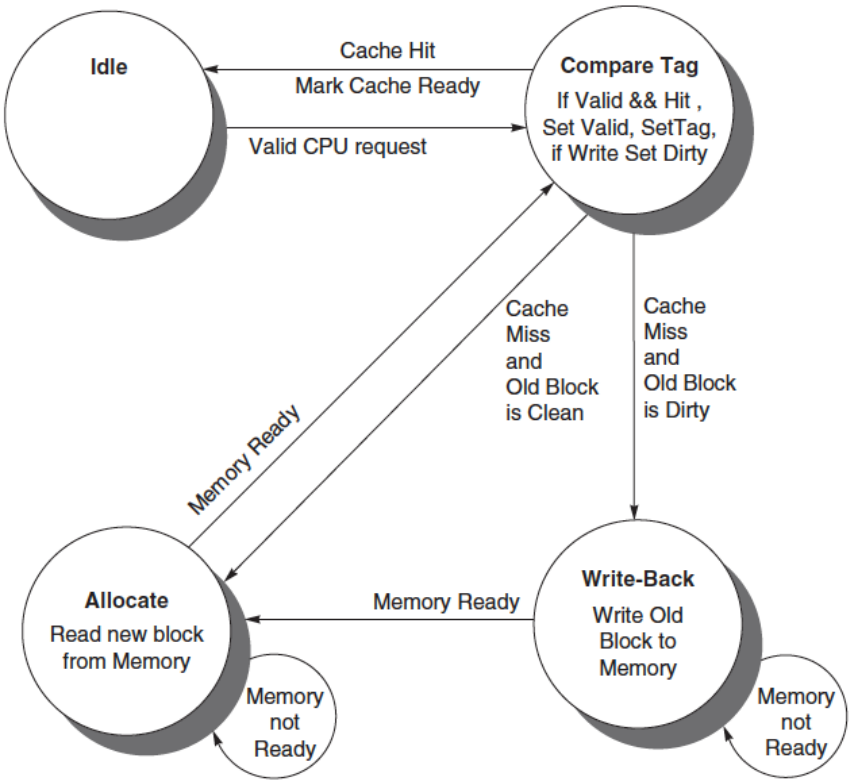
Implementación de un autómata



Máquina de estados finita (Moore)



Controlador simple



Una traza de accesos y sus estados

		TAG (18)				INDEX (10)			2 2		0	1	2	3
0		0	A	B	C	3	4	5	C		✓	✓		✓
0		0	A	B	C	3	4	6	0		✓	✓		✓
0		0	A	B	C	3	4	6	4		✓	✓		
3		0	D	A	A	F	0	0	0		✓	✓		✓
0		0	A	B	C	3	4	6	8		✓	✓		
0		0	A	B	C	3	4	6	C		✓	✓		
0		0	A	B	C	3	4	7	0		✓	✓		✓
2		0	D	A	A	7	0	0	F		✓	✓	✓	✓

¿Qué vimos?

- Cuatro preguntas a la jerarquía
- Las tres C´s
  - Origen de los miss
- Controlador de Cache
  - Señales
  - Máquina de estados