

Tema 4. Memoria Principal

ESTRUCTURA DE COMPUTADORES

Grupo de Arquitectura de Computadores (GAC)



Table of Content

1 Objetivos del Tema

2 Introducción

3 La jerarquía de memoria

- Estructura
- Principio de localidad
- Definiciones

4 La memoria principal

- Conceptos básicos
- Composición interna
- Memoria entrelazada

5 Conclusiones

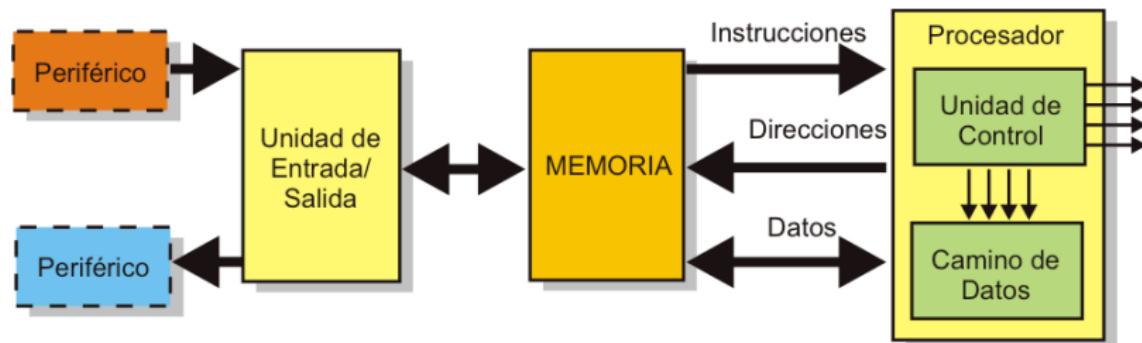
Objetivos del Tema

En este tema se estudiará

- Conceptos básicos del subsistema de memoria de un computador
- Estructura y funcionamiento de la jerarquía de memoria
- Estructura y funcionamiento de la memoria principal

Introducción

El papel de la memoria

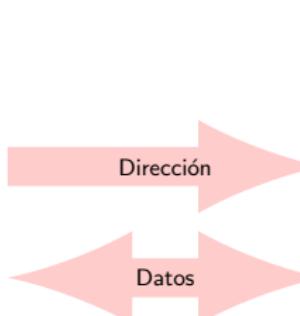


- Es uno de los bloques funcionales básicos, junto con la unidad de control, la unidad de procesamiento y la entrada/salida
- En la memoria se almacenan tanto datos como instrucciones
- Existe un continuo movimiento de información entre la memoria y el procesador. El rendimiento depende en gran medida de ambos elementos.

Introducción

La estructura de la memoria es un conjunto de posiciones, cada una de las cuales contiene un número determinado de bits

- La información se almacena en forma de bits
 - ▶ Los bits se agrupan para formar bytes y palabras
- **Palabra:** Unidad natural de organización de la memoria. Su tamaño suele coincidir con el número de bits utilizados para representar números y con la longitud de las instrucciones

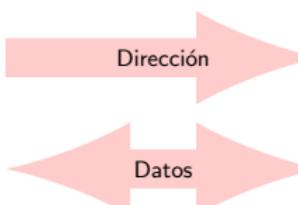


The diagram illustrates the organization of memory. On the left, two pink arrows point towards a table. The top arrow is labeled "Dirección" (Address) and points to the first column of the table. The bottom arrow is labeled "Datos" (Data) and points to the second column of the table.

	Dirección	Contenido
0	0000	0110 0100
1	0001	1011 0101
2	0010	1110 0100
3	0011	0101 0100
4	0100	0010 1110
5	0101	0011 0100
...

Introducción

- Debemos distinguir entre la **dirección** de un dato y su **contenido**
- Una dirección de memoria puede indicarnos la ubicación de un solo **bit**, de un **byte** o de una **palabra** completa
- Con n bits podemos direccionar $m = 2^n$ datos diferentes.
 - ▶ Asimismo, el número de bits necesarios para direccionar m datos diferentes es $n = \log_2(m)$



	Dirección	Contenido
0	0000	0110 0100
1	0001	1011 0101
2	0010	1110 0100
3	0011	0101 0100
4	0100	0010 1110
5	0101	0011 0100
...

Introducción

Características deseables de la memoria:

- Rapidez (bajo tiempo de acceso)
- Alta capacidad
- Coste bajo (coste por bit)

Problema para conjugar estas 3 virtudes:

- A mayor rapidez mayor coste por bit
- A mayor capacidad menor rapidez
- A mayor capacidad menor coste por bit

En general, si tenemos 2 de las 3 características, la otra se disparará

Table of Content

1 Objetivos del Tema

2 Introducción

3 La jerarquía de memoria

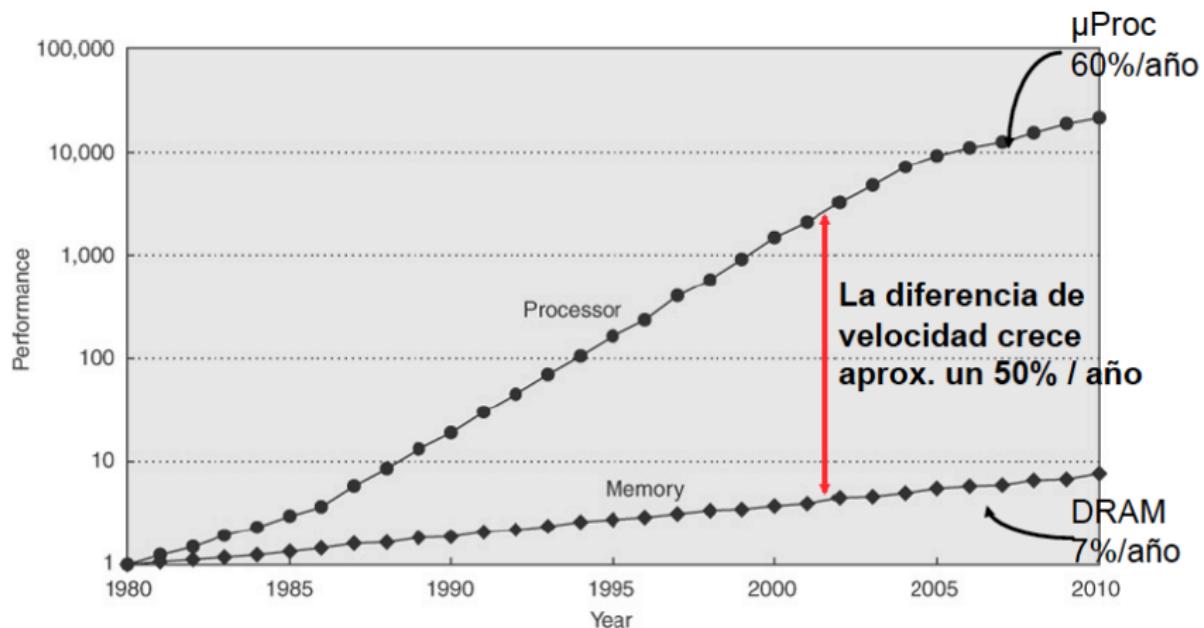
- Estructura
- Principio de localidad
- Definiciones

4 La memoria principal

- Conceptos básicos
- Composición interna
- Memoria entrelazada

5 Conclusiones

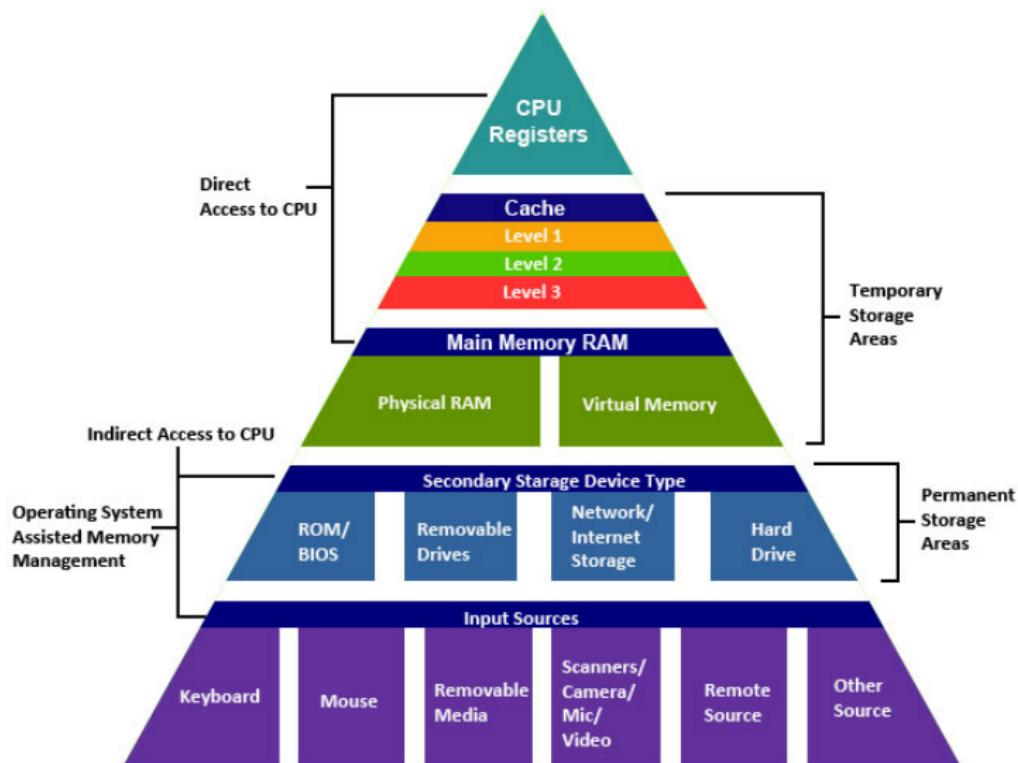
La jerarquía de memoria: Efecto de la ley de Moore



La jerarquía de memoria: Idea básica

- La diferencia entre el rendimiento del procesador y de la memoria aumenta en cada generación debido a la Ley de Moore
 - ▶ El rendimiento del procesador se degrada debido a las continuas paradas que tiene que hacer para esperar por la carga de datos
- La jerarquía de memoria es una organización del subsistema de memoria que trata de mitigar este problema
 - ▶ El subsistema de memoria está compuesto por varios dispositivos de memoria que se organizan jerárquicamente
 - ▶ Memorias rápidas y pequeñas en los niveles superiores de la jerarquía se combinan con otras más lentas y grandes en los niveles inferiores
- El objetivo es crear la ilusión de que toda la memoria es tan rápida como la del nivel superior y tan grande como la del nivel inferior

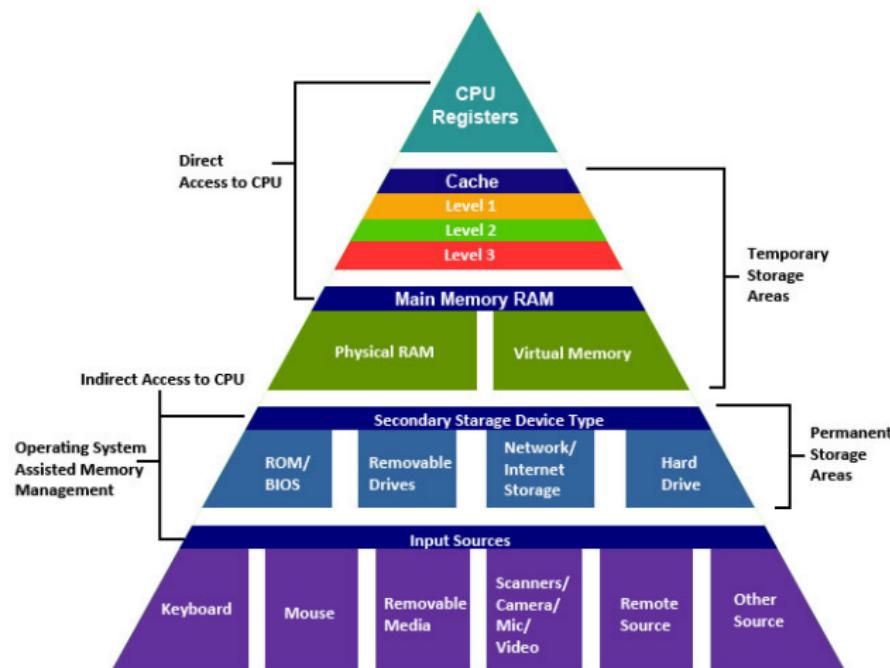
La jerarquía de memoria: Estructura



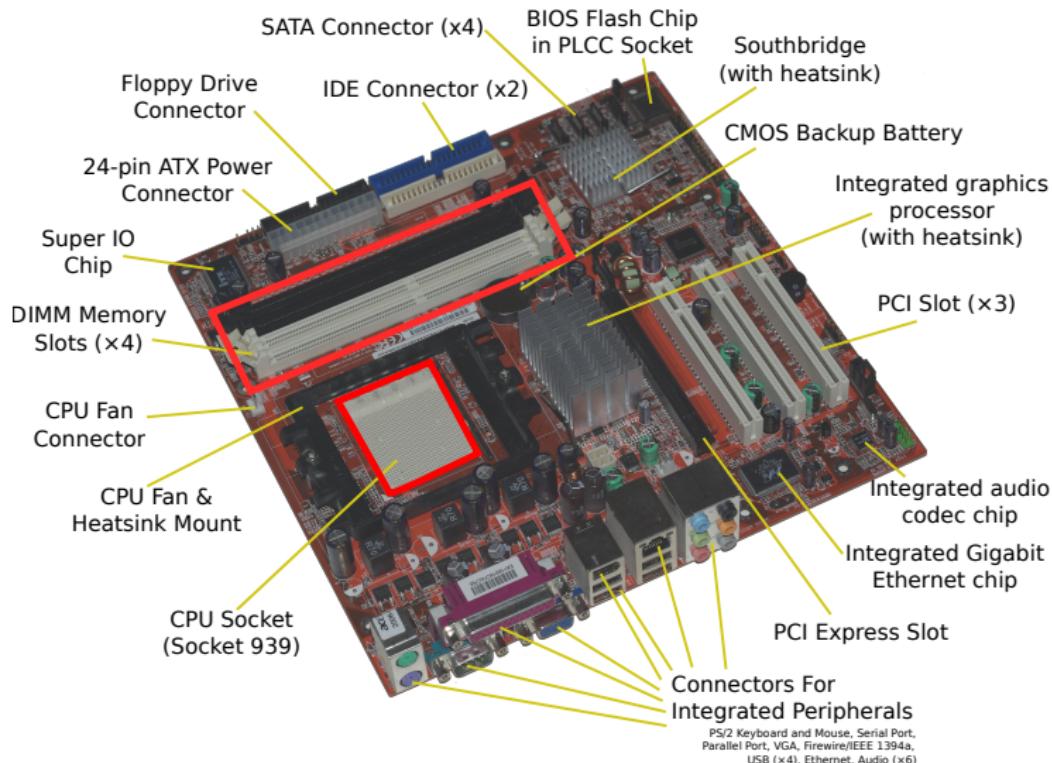
La jerarquía de memoria: Estructura

A medida que descendemos en la jerarquía de memoria

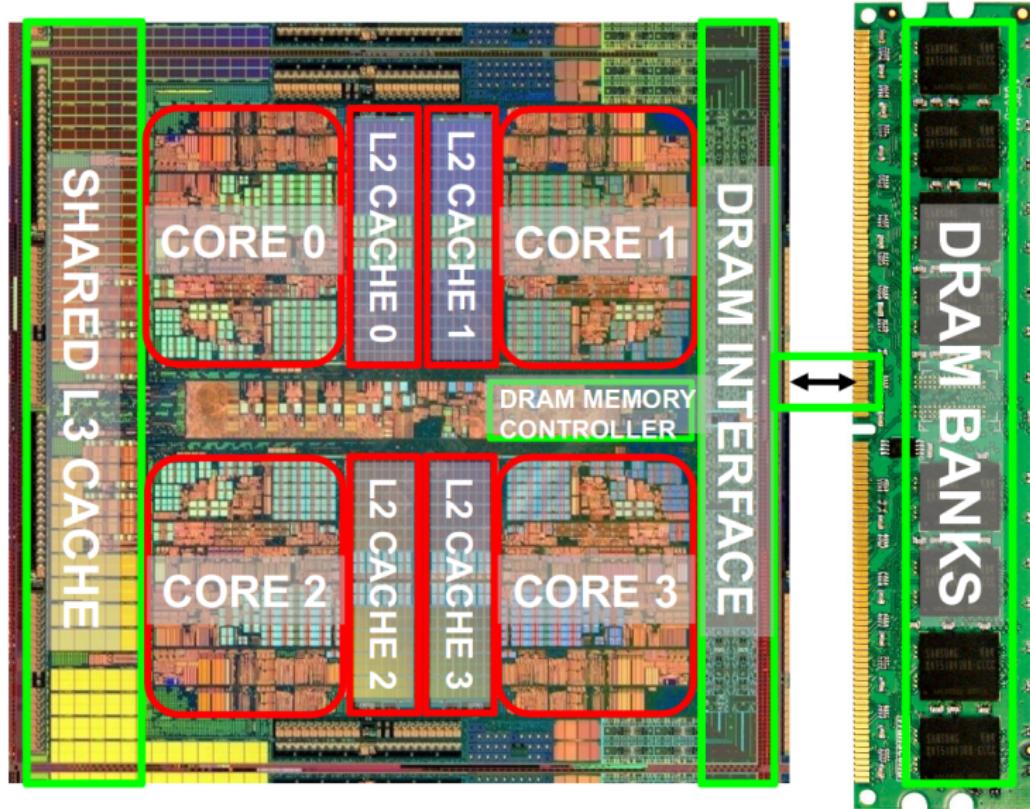
- La memoria se hace más lenta
- Disminuye el coste por bit
- Aumenta la capacidad



Jerarquía de Memoria en la Placa Madre



Jerarquía de memoria (Vista CPU)



(cc) Onur Mutlu, Yoongu Kim

SRAM, DRAM, SDRAM

- Los niveles de la jerarquía de memoria son de acceso aleatorio
 - ▶ La latencia es (aproximadamente) la misma para acceder a cualquier posición de la memoria
- SRAM (Static Random Access Memory)
 - ▶ Acceso muy rápido
 - ▶ No necesita refresco
 - ▶ Mayor coste
 - ▶ Preferible registros y cachés
- DRAM (Dynamic Random Access Memory)
 - ▶ Mucha más capacidad
 - ▶ Menor coste
 - ▶ Preferible para chips independientes
- SDRAM (Synchronous DRAM)
 - ▶ Coordinada por una señal de reloj externa
 - ▶ Utilizada actualmente para memoria principal

La jerarquía de memoria: Funcionamiento

- Cuando se enciende un computador solo los niveles inferiores, los no volatiles, contienen la información.
- A medida que el procesador opera van emitiendo de las direcciones de memoria donde se encuentran tanto las instrucciones como los datos que necesita.
- A medida que se van referenciando datos desde el procesador estos, a parte de ser proporcionados al procesador, se van cargando en los niveles superiores de la jerarquía.
- La carga de los datos se produce en bloques. Es decir, se carga el dato pedido y una cierta cantidad de datos adyacentes a él.

La jerarquía de memoria: Funcionamiento

- Cuando un dato es referenciado, primero se comprueba si está en el nivel superior de la jerarquía.
 - ▶ Si se encuentra se le proporciona al procesador y sino se va a buscar al nivel inmediatamente inferior de la jerarquía.
 - ▶ El proceso desciende de forma recursiva en la jerarquía hasta hallar el dato buscado
 - ▶ Cada vez que se referencia un dato, una vez encontrado se carga en todos los niveles de la jerarquía
- La jerarquía de memoria funciona bien cuando casi todas los bloques de memoria referenciados se encuentran en los niveles superiores de la jerarquía
 - ▶ La clave del éxito de la jerarquía se basa en el **principio de localidad**

La jerarquía de memoria: Principio de localidad

Principio de Localidad

Un programa suele acceder a una parte reducida del espacio de direcciones. Las referencias a memoria generadas por un programa suelen estar agrupadas.

- **Localidad espacial:** Si un dato es referenciado, los datos próximos a él es probable que sean referenciados pronto.
- **Localidad temporal:** Si un dato es referenciado, es probable que ese mismo dato vuelva a ser referenciado pronto.

La jerarquía de memoria: Definiciones

- **Bloque:** es la unidad mínima de transferencia de datos entre los distintos niveles de la jerarquía
- **Acierto:** Cuando un dato se encuentra en un determinado nivel de la jerarquía decimos que su búsqueda terminó en un acierto en ese nivel de la jerarquía.
 - ▶ **Tasa de aciertos:** es la fracción de accesos a un nivel de la jerarquía que resultó en un acierto
- **Fallo:** Cuando un dato no se encuentra en un determinado nivel de la jerarquía
 - ▶ **Tasa de fallos:** fracción de accesos a un nivel de la jerarquía que resultó en un fallo (1 - tasa de aciertos)

La jerarquía de memoria: Definiciones

- **Tiempo de acierto:** tiempo de acceso al nivel superior de la jerarquía
- **Penalización por fallo:** tiempo **adicional** requerido ante un fallo caché, para ir a buscar el dato en los niveles inferiores de la jerarquía e intercambiar un bloque del nivel superior por el bloque buscado y luego proporcionar el dato al procesador

Table of Content

1 Objetivos del Tema

2 Introducción

3 La jerarquía de memoria

- Estructura
- Principio de localidad
- Definiciones

4 La memoria principal

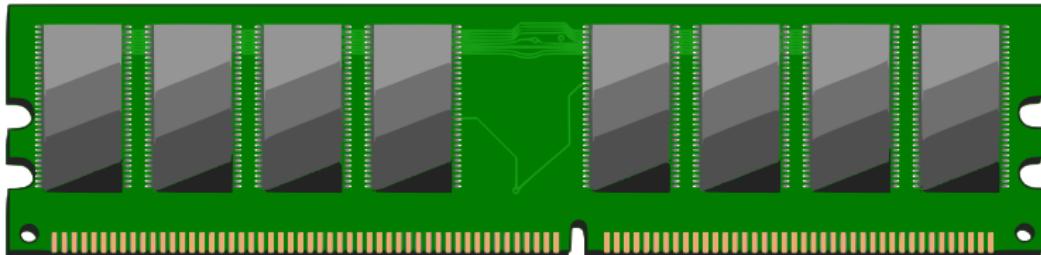
- Conceptos básicos
- Composición interna
- Memoria entrelazada

5 Conclusiones

Memoria principal: Conceptos básicos

- Es el nivel de la jerarquía de memoria ubicado entre la caché y el almacenamiento secundario
- También sirve como interfaz con el subsistema de E/S
- Tecnología que se utiliza:
 - ▶ Generalmente usa SDRAM
 - ▶ Memoria volátil: Cuando se apaga el computador se pierde la información
 - ▶ Memoria basada en tecnología semiconductora: Cada celda de esta memoria está compuesta de uno o varios transistores cuya misión es almacenar un bit de información.

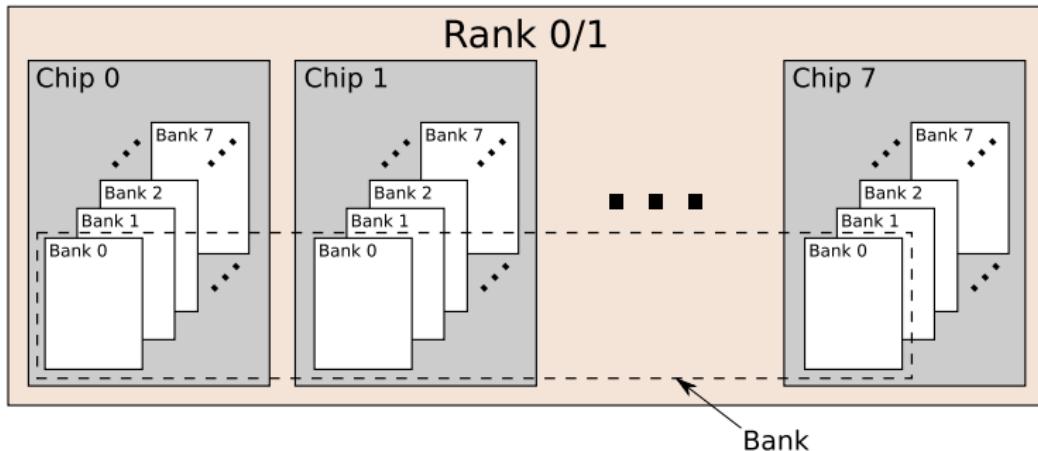
Más conceptos sobre SDRAM



- DIMM (Dual In-line Memory Module)
 - ▶ Es un módulo de RAM tal como lo conocemos hoy en día
 - ▶ Contiene 2 rangos (**uno por cada cara**) de 8 chips
- DDR (Double Data Rate)
 - ▶ Capaz de transferir datos en ambos flancos de reloj (positivo y negativo)
 - ▶ Cada generación soporta mayor frecuencia de reloj, mejora el ancho de banda, la tasa de transferencia, es más eficiente, etc.
 - ▶ DDR (2000), DDR2 (2003), DDR3 (2007), DDR4 (2014)

Descomposición SDRAM

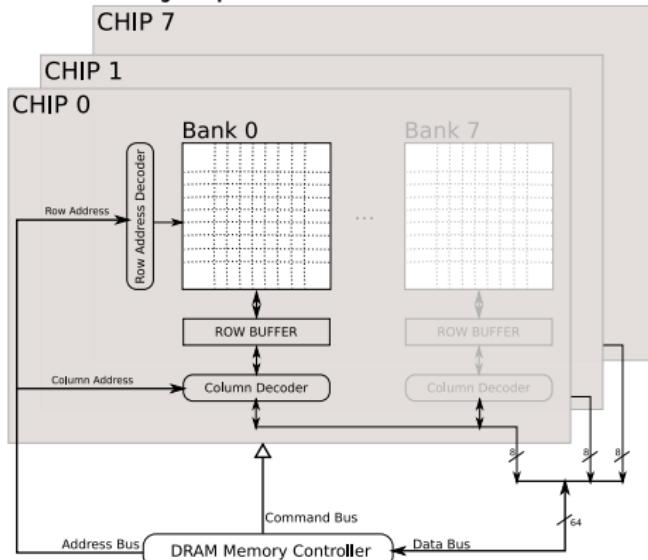
DIMM



- Un DIMM contiene 2 rangos (ranks), uno por cada lado
- Cada rango contiene 8 chips
- Cada chip se compone de 8 bancos de memoria apilados
- El término *banco* referencia también al conjunto de los 8 bancos análogos (con mismo ID) de un mismo rango.

Descomposición SDRAM

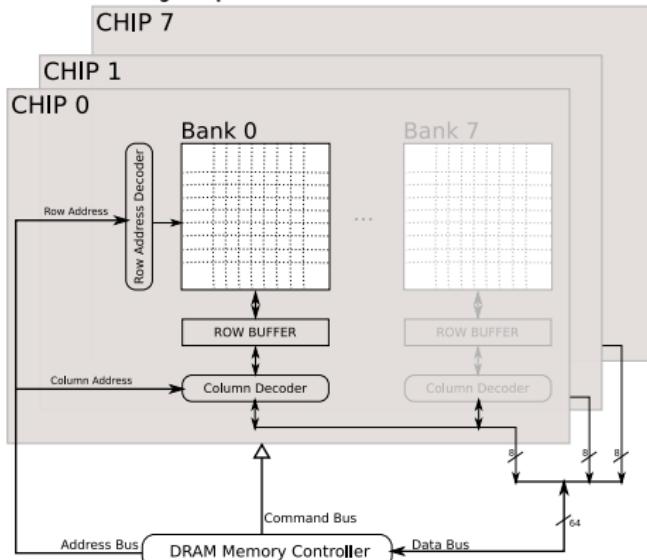
Ejemplo DIMM de 64 bits



- Cada banco contiene una matriz de celdas
- Cada celda almacena un único bit de información
- Podemos buscar la misma celda en el mismo banco en varios chips a la vez
- Permite acceder de forma simultánea a los bits que conforma una palabra
 - ▶ Si accedemos a los 8 bancos en los 8 chips, obtenemos 64 bits

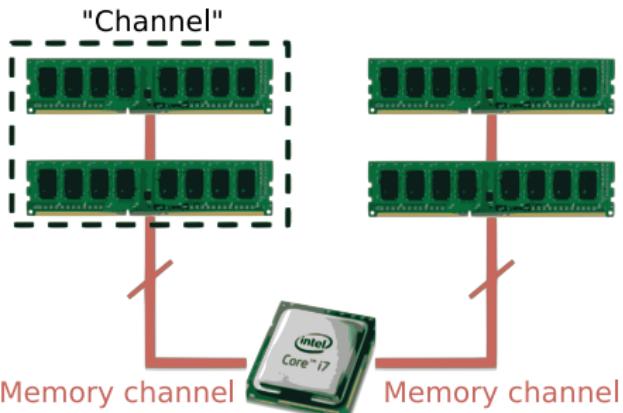
Descomposición SDRAM

Ejemplo DIMM de 64 bits



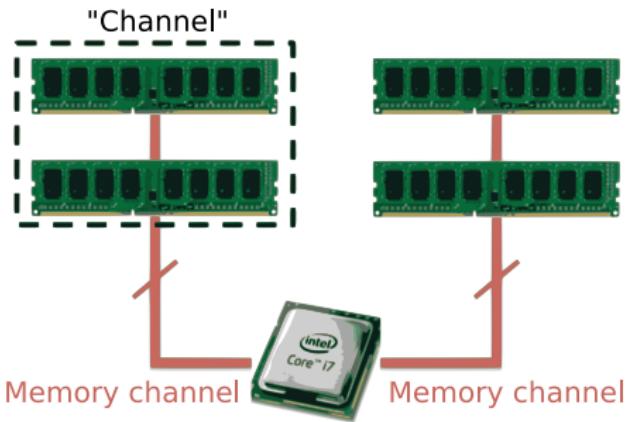
- Para acceder de este modo, la información se almacena **entrelazada** de forma **cíclica**
 - A nivel de bit entre chips
 - A nivel de byte entre bancos

Memoria principal: Canal



- Por encima del acceso paralelo anterior, la placa madre puede tener varios **canales** para la memoria
- Podemos transferir datos de módulos de diferentes canales simultáneamente
 - ▶ Dual-Channel → $\times 2$ Ancho de Banda (imagen superior)
 - ▶ Triple-Channel → $\times 3$ Ancho de Banda
 - ▶ ...
- El acceso a módulos de un mismo canal puede hacerse simultáneamente, pero no la transferencia

Memoria principal: Canal



- Sin embargo, el doble de ancho de banda **no** se traduce en el doble de velocidad de nuestra computadora
- Empíricamente observamos un incremento de velocidad de 15 – 30 %
 - ▶ Los datos se transfieren en bloques (líneas caché)
 - ▶ La latencia sigue siendo alta comparada con niveles superiores de la jerarquía
 - ▶ A pesar de la localidad, no todos los datos transferidos son utilizados

Memoria principal: Rendimiento

- El rendimiento de la memoria principal determina el tiempo que tarda en resolverse un fallo caché
- Parámetros que determinan el rendimiento de la memoria principal
 - ▶ Latencia: o tiempo en acceder al primer dato (depende de la tecnología empleada)
 - ▶ Tiempo de transferencia: tiempo necesario para transferir el resto de la información requerida (depende del ancho de banda entre la caché y la memoria principal)

Memoria principal: Rendimiento

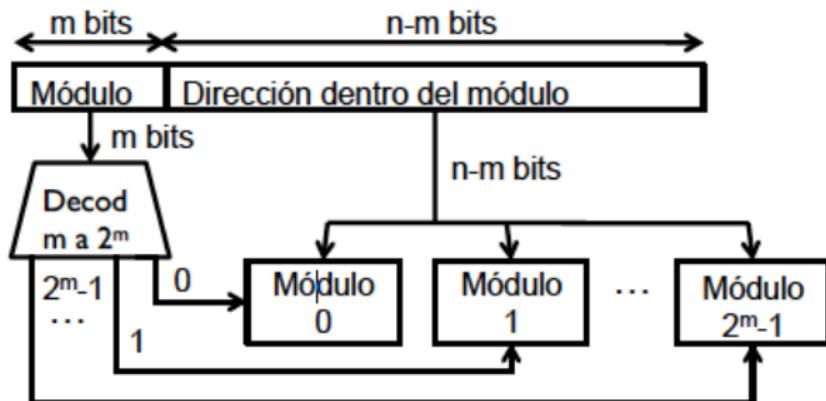
- Formas de aumentar el ancho de banda entre la caché y la MP:
 - ▶ Aumentar el ancho de la memoria.
 - ★ El bus y la memoria pasan a tener varias palabras de ancho
 - ★ Se utiliza un multiplexor para seleccionar qué palabra se le proporciona a la caché en cada momento
 - ▶ Utilizar bancos de memoria independientes
 - ★ La opción más sencilla pero también la más costosa
 - ▶ Utilizar memoria entrelazada

Memorias entrelazadas

La información está repartida entre varios módulos de memoria. Así se pueden atender varias peticiones de forma solapada/simultánea.

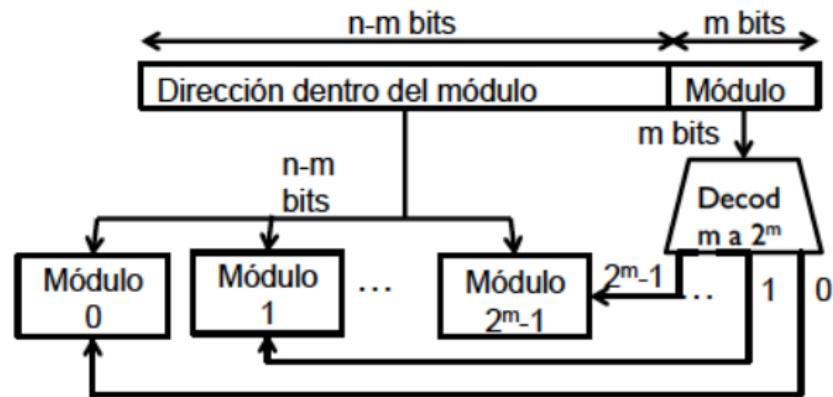
- Se sigue pagando el coste de transmitir cada palabra secuencialmente dentro de un mismo canal, pero se evita pagar más de una vez la latencia de acceso
- Se hacen más rápidas las escrituras ⇒ importante en escritura directa
- Funcionan al mismo tiempo todos los módulos ⇒ más consumo
- Esquemas de entrelazado:
 - ▶ Entrelazado de orden superior
 - ▶ Entrelazado de orden inferior

Entrelazado de orden superior (consecutivo)



- Memoria total: $N = 2^n \Rightarrow$ Dirección física: n bits
- Número de módulos: $M = 2^m$
- Al i -ésimo módulo le corresponden las direcciones consecutivas $i \times 2^{n-m}$ a $(i + 1) \times 2^{n-m} - 1$
- Los m bits más significativos identifican el módulo y el resto un *desplazamiento* dentro del módulo

Entrelazado de orden inferior (cíclico)



- Memoria total: $N = 2^n \Rightarrow$ Dirección física: n bits
- Número de módulos: $M = 2^m$
- Al i -ésimo módulo le corresponden las direcciones de la forma $k \times M + i$ con $k = 0, 1, 2, \dots, 2^{n-m} - 1$
- Los m bits menos significativos identifican el módulo y el resto un *desplazamiento* dentro del módulo

Entrelazado: conflictos de memoria

- Con cualquiera de los dos esquemas se pueden obtener M palabras en paralelo por cada acceso a memoria
- **Conflictos de memoria:** varias direcciones requieren simultáneamente el acceso a un mismo módulo
- Los conflictos de memoria son mayores en orden superior debido a la secuencialidad
- En sistemas multiprocesador es a veces mejor el entrelazado de orden superior cuando las tareas son disjuntas o interaccionan poco entre sí (lo cual no siempre es cierto)
- Se suele utilizar el entrelazado de orden inferior
- Ventajas del superior:
 - ▶ Expansibilidad
 - ▶ Fiabilidad: un fallo se restringe a un área localizada del espacio de direcciones

Table of Content

1 Objetivos del Tema

2 Introducción

3 La jerarquía de memoria

- Estructura
- Principio de localidad
- Definiciones

4 La memoria principal

- Conceptos básicos
- Composición interna
- Memoria entrelazada

5 Conclusiones

Conclusiones

- La jerarquía de memoria trata de mitigar la enorme diferencia de rendimiento entre el procesador y la memoria
- El rendimiento de la memoria principal es crucial para el rendimiento de la jerarquía de memoria
- El entrelazamiento en la memoria permite mejorar el rendimiento sin incrementar demasiado el coste