Tema-3-parte-2-Sistemas-de-memor...

atmuntenas

Fundamentos de Informática

1º Grado en Ingeniería de las Tecnologías de Telecomunicación

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación
Universidad de Granada

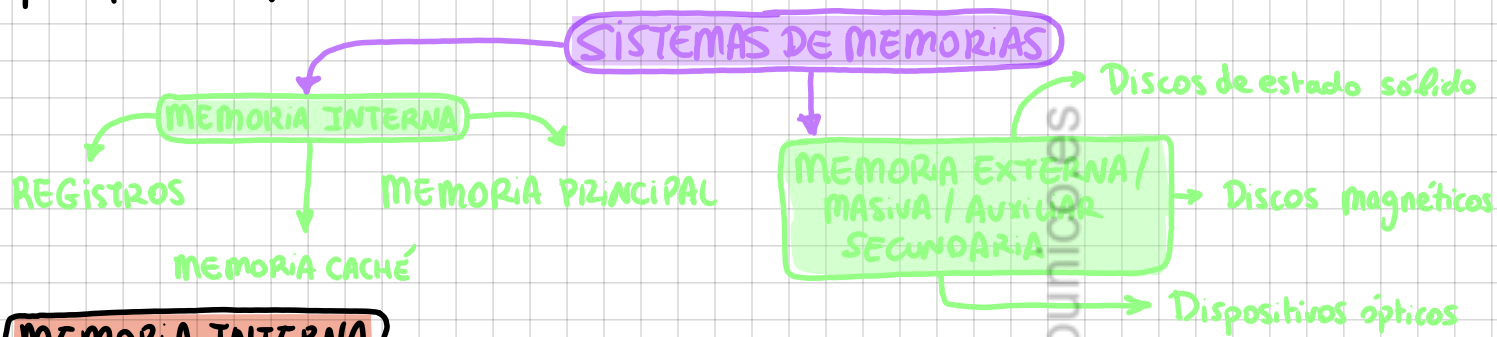
Accede al documento original

Tema 3 - Sistemas de memorias y Conexiones

SISTEMAS DE MEMORIA

El sistema de memoria está formado por las unidades donde se almacenan los datos y las instrucciones. Sirven para trabajar y almacenar con información.

Tradicionalmente se consideran dos estructuras básicas de memoria diferenciadas principalmente por su velocidad y volatilidad:



MEMORIA INTERNA

En esta sección estudiaremos las unidades de memorias internas: los registros, la memoria caché o semiconductora y la memoria principal. La primera ya la estudiamos con anterioridad.

Existen dos tipos de memorias semiconductoras:

→ las RAM estáticas (SRAM) en las que cada celda de memoria está constituida por un biestable de unos 6 transistores. ↳ Static Random Access Memory

→ las RAM dinámicas (DRAM), en el que cada celda de memoria consta de un solo transistor ↳ Dynamic Random Access Memory

Las SRAM son más rápidas que la DRAM, pero la primera es más compleja y menos miniaturizables que las DRAM, teniendo cada chip mucha menos capacidad de almacenamiento.

Existen diversos parámetros para determinar las características de la memoria, como:

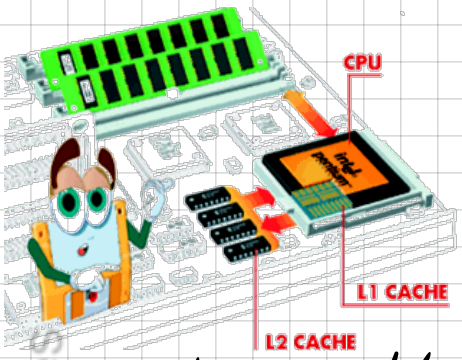
→ Tiempo de acceso a memoria t_a o latencia: tiempo que transcurre, desde que se presenta una dirección de memoria, hasta que el dato queda memorizado o está disponible para su uso.

→ Tiempo de ciclo de memoria t_c: tiempo mínimo que debe transcurrir entre dos accesos

sucesivos.

→ Ancho de banda: número de bytes que se pueden transmitir por segundo entre la memoria y el procesador. Depende del tiempo de acceso a memoria, el número de bytes a los que se puede acceder en paralelo y la capacidad de transferencia del bus de interconexión.

MEMORIA CACHÉ



La memoria principal es mucho más lenta que el procesador, lo que ralentiza el acceso de datos. Para solucionar esto, se utiliza la memoria caché, que es más rápida y se encuentra entre el procesador y la memoria principal. La caché está formada por SRAM, de capacidad menor y es más cara.

La memoria aprovecha los principios de localidad de referencias: localidad espacial (carga no solo el dato solicitado, sino también los datos próximos, bloques de datos) y localidad temporal (mantiene los datos en caché por si necesitan de nuevo).

Si el procesador no encuentra los datos en la caché, se produce un fallo en caché, y se cargan bloques completos de datos en la línea de caché. Si los datos están presentes, ocurre un acierto de caché y se utilizan directamente.

MEMORIA PRINCIPAL

La memoria principal está organizada en palabras de memoria, que son conjuntos de bits que se pueden leer o escribir a la vez. Cada palabra tiene un ancho de palabra que puede ser de 8, 16, 32 o 64 bits. Para acceder a las posiciones de memoria se usan un bus de direcciones y uno o más buses de datos. La capacidad máxima de la memoria principal en bytes se calcula con la siguiente fórmula:

$$C_{mp} = 2^m \times \frac{n}{8}$$

n es el ancho de palabra
m es el número de hilos del bus de direcciones.

Los módulos de memoria se construyen con circuitos DRAM y se agrupan en módulos como SIMM (Single In-line Memory Module), DIMM (Dual In-line Memory Module), SODIMM (Small Outline - DIMM) y RIMM (Rambus In-line Memory Module).

En cuanto al direccionamiento de palabra, se pueden usar dos métodos:

→ Memorias direccionables por palabras, donde a cada dirección corresponde a una palabra completa. Ejemplo: En una memoria de 32 bits, las direcciones estarán alineadas a múltiplos de 4 bytes (32 bits).

→ Memorias direccionables por bytes, donde cada dirección corresponde a un byte de memoria.

En este último caso, las direcciones de las palabras no siempre estarán alineadas a múltiplos de su tamaño, lo que se llama direccionamiento no lineado. Para asegurar direcciones alineadas, los bits menos significativos de la dirección deben ser 0, dependiendo del tamaño de la palabra.

Los accesos de memoria pueden realizarse de tres formas:

→ Por palabras: que es el acceso normal

→ Por bytes: donde se lee y escribe solo el byte solicitado

→ Por ráfagas: donde se transfieren bloques de datos consecutivos, como al cargar un programa o guardar un archivo. En este caso, solo se indica al sistema de memoria solo la dirección inicial del bloque y su tamaño.

Para mejorar la velocidad de acceso, se usa el entrelazado de memoria, que distribuye los datos en módulos de memoria en paralelo. Así, las posiciones consecutivas de memoria se distribuyen entre módulos, y al acceder a una palabra, se acceden a varias palabras consecutivas. El direccionamiento se organiza de tal manera que los primeros bits seleccionan la palabra dentro de cada módulo y los últimos bits seleccionan el módulo que contiene la palabra solicitada.

MEMORIAS EXTERNAS

En este apartado se presentan una serie de dispositivos basados en principios magnéticos y ópticos que actúan como periféricos extendiendo la memoria principal. El conjunto de estos dispositivos constituye la memoria externa del computador, ayudando a resolver el problema de la volatilidad y la capacidad limitada de la memoria interna.

Cuando se apaga el comput. la información se borra.

Los principales soportes utilizados son:

→ Memorias magnéticas: disco y cinta magnética
→ Memorias ópticas: CD-ROM y DVD-ROM
→ Memorias magneto-ópticas
→ Memorias flash: SSD y "USB".

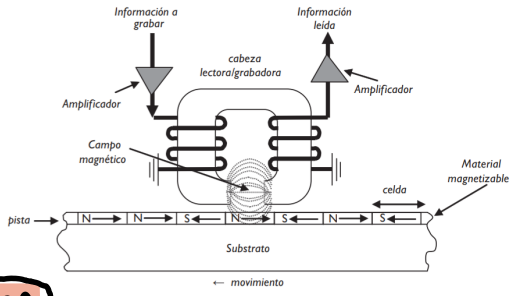


Figura 10.7. *Esquema que muestra el fundamento de la grabación y lectura en un soporte magnético por una cabeza lectora/grabadora.*

ESCRITURA Y LECTURA DE INFORMACIÓN EN FORMA MAGNÉTICA

Los discos y cintas magnéticas consisten en un sustrato de plástico o aluminio recubierto con un material magnetizable, como óxido férrico o de cromo, donde la información se almacena en **celdas** dispuestas en **pistas**. Cada celda puede tener un campo magnético orientado en dos direcciones posibles (norte/sur), lo que permite representar los valores binarios de 0 ó 1.

Para leer/escribir información, se utiliza una **cabeza de lectura/escritura** que, mediante una bobina, genera un campo magnético para escribir celdas o detecta variaciones en el campo magnético para leer los datos.

En 1990, IBM desarrolló las **cápsulas magnetorresistivas**, empleadas para mejorar la lectura de datos. Esta cabeza aprovecha los cambios manifestados en la variación de la resistencia eléctrica bajo el efecto de los campos magnéticos de las celdas, amplificando dicha alteración para convertirlo en un valor digital.

La información no se transmite en **bloques** o **registros físicos**, lo que significa que la lectura o escritura se realiza en unidades más grandes de datos. Este acceso a los datos viene determinado por el **tiempo de acceso medio**, que es el promedio de tiempo necesario para acceder a una celda o bloque de datos, y puede variar dependiendo de la posición de la cabeza lectora en el disco o cinta.

Los dispositivos de **acceso secuencial** (ej: cintas magnéticas), requieren que la cabeza lectora lea los datos en un orden específico, ralentizando el acceso. En cambio, los dispositivos de **acceso directo** permiten mover la cabeza directamente al bloque de datos deseado, haciendo el acceso más rápido.
→ ej: discos magnéticos.

DISCOS MAGNÉTICOS

Son el principal soporte de memoria principal de almacenamiento de memoria auxiliar, debido a su capacidad de acceso directo.

La información se graba en pistas concéntricas de un disco circular recubierto con material magnetizable, con cada pista dividida en sectores, que son las unidades de almacenamiento de la información, cada uno con una capacidad típica de 512 bytes.

El acceso se realiza en bloques o registros físicos, y el tiempo necesario para acceder a un bloque depende de tres factores clave.

→ Tiempo de búsqueda (T_b) es el tiempo necesario para posicionar la cabeza lectora sobre la pista correcta. Depende de la distancia entre la cabeza y la pista objetivo y la rapidez con que el motor mueve la cabeza.

T_b = tiempo reacción motor + nº pistas a recorrer × tiempo medio para atravesar una pista

→ Tiempo de lectura/escritura (T_r) es el tiempo necesario para leer o escribir los bytes en bloque. Depende de la capacidad de pista y la velocidad de disco.

$$T_r = \frac{\text{Cantidad datos leer/escribir}}{\text{Capacidad de la pista}} \times \text{tiempo medio por pista}$$

→ Tiempo de espera (T_E) es el tiempo que debe esperar la cabeza hasta el sector deseado se posicione bajo a ella debido a la rotación de disco.

$$T_{ESPERA} = \frac{1}{2 \cdot \text{Velocidad disco en rps } (w_r)}$$

El tiempo total de acceso $T_a = T_b + T_E + T_r$

Según la velocidad de giro, distinguimos varios tipos de discos magnéticos:

→ CAV (Constant Angular Velocity): la velocidad de lectura/escritura se mantiene cte., lo que significa en una mayor densidad de grabación en las pistas internas.

→ ZCAV (Zoned CAV): la velocidad de grabación aumenta en las pistas exteriores, lo que permite una mayor capacidad y velocidad de acceso a las pistas externas.

→ CLV (Constant Linear Velocity): la velocidad de rotación varía según la posición de la pista, lo que es típico en los discos ópticos.

En cuanto a la controladora de disco, este contiene un procesador específico (secuenciador) que admite órdenes para controlar tareas como arranque, leer, escribir y formatear. También controla el movimiento del peine (el lector magnético), detecta y corrige errores, y convierte los bytes (que llegan en paralelo) en patrones de grabación (usando codificación RLL).

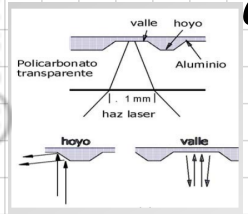
Las órdenes y las transferencias de datos se dan a través de puertos E/S, y la controladora dispone de un gran buffer o caché para permitir transferencias de datos a la CPU o DMA (Direct Memory Access) a velocidades elevadas. La lectura y escritura de la superficie se realizan mediante el uso de la caché, trabajando en cilindros o pistas completas.

UNIDADES RAID

Una unidad RAID (Redundant Array of Independent Disks) o agrupación redundante de discos independientes, es un conjunto de discos que funcionan en paralelo y son considerados por el sistema operativo como única unidad.

El objetivo de este tipo de unidades es doble: aumentar la velocidad y mejorar la seguridad y fiabilidad de los datos almacenados.

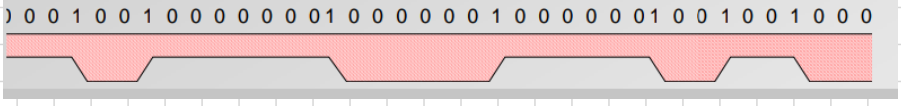
DISCOS ÓPTICOS



Los discos ópticos, como los CD o DVD, son medios de almacen. que utilizan láseres para escribir y leer datos. Velocidad de giro es CLV.

El proceso de escritura implica que un láser cree hoyos microscópicos (pits en la superficie del disco, representando datos binarios. El láser modifica el material foto-sensible, creando una secuencia de pits y áreas planas (lands).

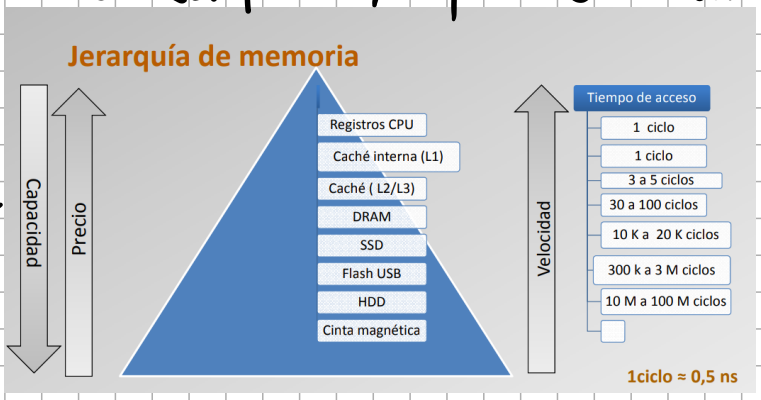
Al leer el disco, otro láser detecta los reflejos de estas áreas, interpretando la información.



JERARQUÍA DE MEMORIA

La forma de almacenamiento de información de un computador, se puede establecer una jerarquía, bajo cuatro puntos de vista.

→ tamaño o capacidad
→ tiempo de acceso, lo menor posible
→ ancho de banda, alto
→ coste por bit, reducido.



La organización del sistema de memoria se basa en una jerarquía de niveles con diferentes propiedades de acceso y velocidad. Los programas y datos están almacenados en disco, pero las instrucciones y datos en ejecución deben estar en los registros del procesador para ser utilizados rápidamente.

Cuando el procesador necesita datos, primero intenta obtenerlos en niveles más rápidos, como la caché. Si la información se encuentra en un nivel, se produce un **acierto**, sino, se produce un **fallo**, y el procesador debe buscar en niveles más bajos.

Dicho esto, podemos definir la **tasa de aciertos** es la probabilidad de que los datos estén en el nivel i deseado:

$$\gamma_{aciertos,\,nivel} = \frac{\text{nº accesos realizados con éxito}}{\text{nº total de accesos al nivel}}$$

Y también la **tasa de fallos** es el opuesto, la probabilidad de no encontrar los datos en el nivel i deseado:

$$\gamma_{fallos,\,nivel} = \frac{\text{nº de fallos}}{\text{nº de accesos}} \Rightarrow \gamma_{aciertos,\,nivel} + \gamma_{fallos,\,nivel} = 1$$

Cuando ocurre un fallo, los datos del nivel inferior son copiados al nivel actual. Si un nivel está lleno, se utiliza un **algoritmo de reemplazo** para decidir qué bloque desalojar.

El **tiempo medio de acceso efectivo** a un nivel de memoria se calcula ponderando el tiempo de acceso del nivel con las tasas de aciertos y fallos.

$$t_{ae,\,nivel} = \gamma_{aciertos,\,nivel} \cdot t_{nivel} + \gamma_{fallos,\,nivel} \cdot t_{ae,\,i+1}$$

(t_{nivel}: tiempo de acceso nivel)

Este cálculo incluye el tiempo de acceso al nivel actual y, en caso de fallo, al nivel inferior.