

FICHEROS Y BASES DE DATOS (E44)

3º INGENIERÍA EN INFORMÁTICA

Tema 1.

Aspectos Básicos de los Ficheros.

- 1.- Jerarquía de Memoria.
- 2.- Ficheros Lógicos y Ficheros Físicos.
- 3.- Acceso a los Datos situados en Ficheros.
- 4.- Coste de Acceso a Dispositivo de Almacenamiento.
 - 4.1. Coste de Acceso a Discos.
 - 4.2. Coste de Acceso a Cintas.

(Capítulos 1, 2 y 3 del Folk)

(Capítulo 4 del Elmasri)

JERARQUÍA DE MEMORIA

Manejo de la Información

- El objetivo básico del manejo de información es la manipulación de la mayor cantidad de información del modo más eficiente posible.
- Sobre los datos se pueden definir diferentes tipos de operaciones, entre las que aparecen:
 - Consulta y búsqueda de información.
 - Actualización de la información, mediante la inserción, el borrado y la modificación.
- Por tanto se desea maximizar las prestaciones de estas operaciones, para lo cual se debe explotar la jerarquía de memoria.

Memoria Primaria

- La primera idea es almacenar la información en el medio más rápido posible, para asegurar una rápida gestión.
- El coste de acceso a memoria primaria, o memoria RAM, es fijo y reducido.
- Así pues, sería interesante el almacenamiento de la información en memoria primaria
- Pero la utilización de este tipo de memoria presenta diferentes problemas:
 - Suele ser un medio bastante caro, por lo que se suele limitar su capacidad.
 - La información se pierde al producirse un fallo de corriente eléctrica.

JERARQUÍA DE MEMORIA

Memoria Secundaria

- Estas limitaciones aconsejan la utilización de la memoria secundaria, formada por diferentes tipos de dispositivos como discos magnéticos, cintas magnéticas y discos ópticos.
- Sus propiedades son las siguientes:
 - Presenta un coste por byte mucho menor que la memoria RAM.
 - Preserva su contenido al producirse un fallo de corriente eléctrica.
 - Es bastante más lenta que la memoria RAM, unas 250.000 veces para discos magnéticos, y además el coste de acceso no es fijo.

Definición de Ficheros

- La información en memoria secundaria se almacena en ficheros, que se define como una Colección de Información Relacionada.
- Es importante destacar, que la información que aparece en un fichero no se encuentra organizada.
- Debido al alto coste temporal del acceso a la memoria secundaria, se desea
 - Maximizar la información recuperada.
 - Minimizar el número de accesos.
- Este es el fundamento para la utilización de estructuras de la información específicas.

JERARQUÍA DE MEMORIA

Manejo de la Jerarquía de Memoria

- La solución más adecuada es explotar las características de ambos tipos de memoria.
- Resulta conveniente almacenar la información en memoria secundaria, ya que presenta una mayor capacidad y preserva su contenido ante cortes de corriente eléctrica.
- Pero cuando se desea procesar, debe de importarse a memoria primaria, para obtener una velocidad adecuada.
- Si la información se modifica debe de volver a exportarse a memoria secundaria.

Manejo de Buffers

- La diferente de velocidad entre los dos tipos de memoria es muy grande, por lo que resulta interesante definir algún tipo de estrategia que reduzca este diferencial.
- Un buffer se define como un conjunto de bytes que son leídos o escritos desde un dispositivo de almacenamiento, en la memoria primaria.
- Cuando se desea leer una información, se lee un bloque de información en el que aparece.
- La modificación de un dato se realiza sobre el buffer, que posteriormente debe ser enviado al dispositivo de almacenamiento.
- La utilización de esta técnica permite reducir el número de accesos a memoria secundaria.

FICHEROS LÓGICOS Y FICHEROS FÍSICOS

Definiciones

- Desde un punto de vista físico, un fichero se define como un conjunto de bytes que se almacenan en memoria secundaria.
- Desde el punto de vista de una aplicación, un fichero es su conexión con el mundo exterior, posibilitándole el envío y la recepción de información.
- De este modo se definen el Fichero Físico y el Fichero Lógico.
- La conexión entre un fichero físico y un fichero lógico es realizada por el sistema operativo, a partir de una orden definida en el programa.
- En esta orden debe aparecer el nombre físico del fichero, y da como resultado el nombre lógico dentro del programa.

`fd = open ("fichero.dat", modo);`

- Dicha conexión se rompe mediante otra orden de sistema operativo, que indica que el nombre lógico ya no corresponde con el nombre físico anterior.

`close (fd);`

Los buffers asociados que hayan sido escritos se envían al dispositivo de almacenamiento, para que éste los almacene definitivamente.

- Con posterioridad, el mismo nombre lógico puede ser utilizado para otro fichero físico, y el fichero físico original puede ser conectado a otro nombre lógico.

ACCESO A LOS DATOS EN FICHEROS

Panorama General

- El tráfico de la información desde y hacia un fichero involucra al sistema operativo y a una serie de dispositivos concretos.
- El proceso se inicia con la aplicación de una operación de acceso o modificación sobre el fichero lógico.
- Esta orden es transmitida al sistema operativo que se encarga de asegurar su completa finalización.

Administrador de Ficheros

- El administrador de ficheros es la parte del sistema operativo que se encarga de la gestión de los ficheros.
- Su primera tarea es comprobar que existe una conexión entre el fichero lógico y un fichero físico determinado.
- Seguidamente se define en que parte del fichero se desea realizar la operación, y si ésta se encuentra en un buffer de memoria.
- En caso negativo, será necesario leer la información sobre un buffer de memoria.
- Si aparece la información en un buffer, la operación se puede completar sobre éste.
- Si la operación modifica el buffer, éste deberá ser enviado al dispositivo de almacenamiento para su actualización.

ACCESO A LOS DATOS EN FICHEROS

Procesador de Entrada/Salida

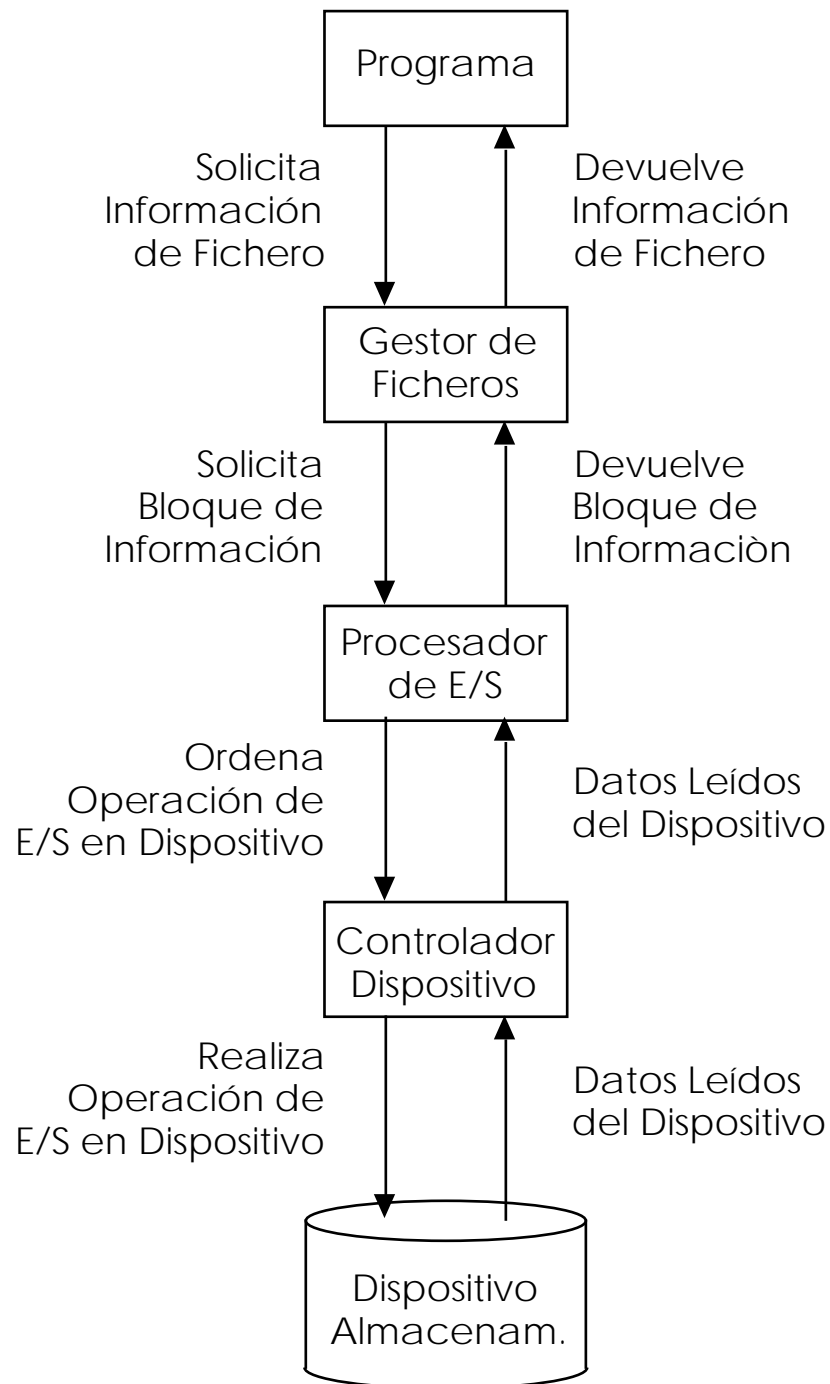
- El procesador de entrada/salida se encarga de controlar el tráfico de información desde y hacia la memoria primaria.
- Este dispositivo tiene un funcionamiento autónomo, liberando al procesador de esta costosa tarea.
- Actúa sobre diferentes tipos de dispositivo de almacenamiento, a partir de las órdenes recibidas desde el sistema operativo.
- Por lo tanto, sólo se encarga de preparar la información para que sea procesada por los dispositivos de almacenamiento.

Controladores de Dispositivos

- Este dispositivo realiza la operación de modo explícito, a partir de la petición realizada por el procesador de entrada/salida.
- Para ello, debe notificar al hardware del dispositivo las operaciones que debe realizar para completar la operación.
- En una lectura, la información fluye desde el dispositivo hasta el controlador, el procesador de entrada/salida y por último a la memoria.
- Una vez localizado el buffer en memoria se procede a realizar la operación en concreto.
- Una escritura debería repetir el proceso en sentido inverso.

ACCESO A LOS DATOS EN FICHEROS

Diagrama de Lectura de Información



ACCESO A LOS DATOS EN FICHEROS

Número de Buffers y Velocidad de Acceso

- El manejo de buffers por parte del administrador de ficheros permite reducir el número de accesos a memoria secundaria.
- Pero una cuestión fundamental es el número de buffers a utilizar.
- Si sólo se utiliza un buffer, un problema que realice lecturas y escrituras de modo alterno, debería leer un bloque en cada operación.
- Esto se resuelve mediante la utilización de un buffer para escritura y otro para lectura.
- Pero la lectura, o escritura, alterna sobre varios ficheros puede provocar el mismo problema.
- Otra alternativa es la utilización de ambos bloques para lecturas y escrituras de modo alternado.
- La generalización de esta idea es el caso real, varios buffers que se manejan de modo indistinto para lecturas y escrituras.
- La gestión de estos buffers es realizada por el administrador de ficheros, aunque el usuario puede controlar el número de buffers.
- Si todos los buffers están ocupados, se debe vaciar uno de ellos para posibilitar una lectura.
- Normalmente se utiliza al algoritmo LRU, es decir, se vacía el buffer menos recientemente utilizado.

ACCESO A LOS DATOS EN FICHEROS

Funcionamiento del Administrador de Ficheros

- Las características del sistema operativo es fundamental para que la utilización de los buffers se realice de modo eficiente.
- En este caso, se considera el modo en el que el administrador de ficheros maneja la información que el programa desea escribir sobre memoria secundaria.
- Esta información no suele estar disponible, por lo que el programador debe preocuparse de conseguirla.
- Una opción es el Modo de Movimiento, que obliga a que los datos sean copiados desde la memoria de la aplicación hacia los buffers y viceversa.
- Esta alternativa suele ser muy costosa, y poco deseable.
- Para resolver este problema se utiliza el Modo de Direcciones, que puede desarrollarse de dos formas.
- La primera permite manejar al administrador de ficheros la memoria de la aplicación.
- En la segunda alternativa el administrador de ficheros suministra la dirección de los buffers al programa, que puede manejarlos de modo directo.

COSTE DE ACCESO A DISPOSITIVOS

Tipos de Dispositivos

- Existen diferentes criterios para clasificar los dispositivos de almacenamiento.
- Uno de los más adecuados se centra en el modo en el que se accede a un dato dentro del medio de almacenamiento.
- Según este criterio, se pueden distinguir dos tipos de dispositivos de almacenamiento:
 - Los Dispositivos de Acceso en Serie, como las cintas magnéticas.
 - Los Dispositivos de Acceso Directo, como los discos magnéticos.
- En los primeros, el acceso a un dato requiere el acceso a todos los datos que le preceden físicamente en el dispositivo.
- Por su parte, los segundos pueden acceder de modo directo a un dato.
- Esta diferencia tiene una gran importancia para la elección de un dispositivo concreto en la resolución de un problema.
- Así, los primeros se utilizan para almacenar información que deba de ser leída o escrita de modo global.
- Mientras que los segundos se utilizan en el manejo de ficheros cuyo criterio de acceso es más aleatorio.

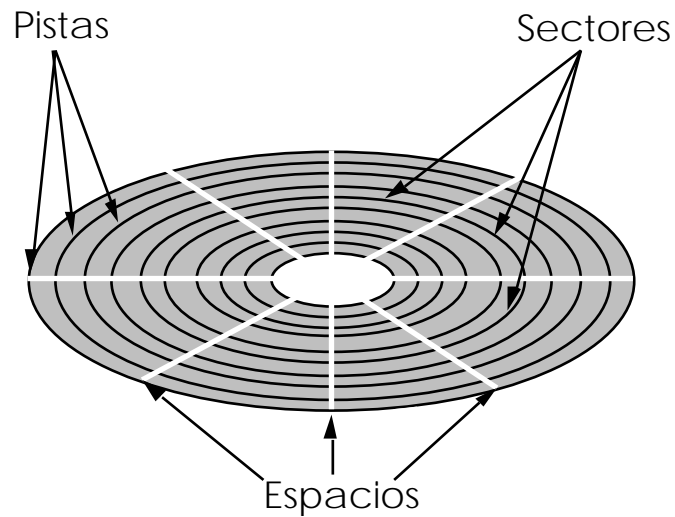
COSTE DE ACCESO A DISCOS

Organización de un Disco

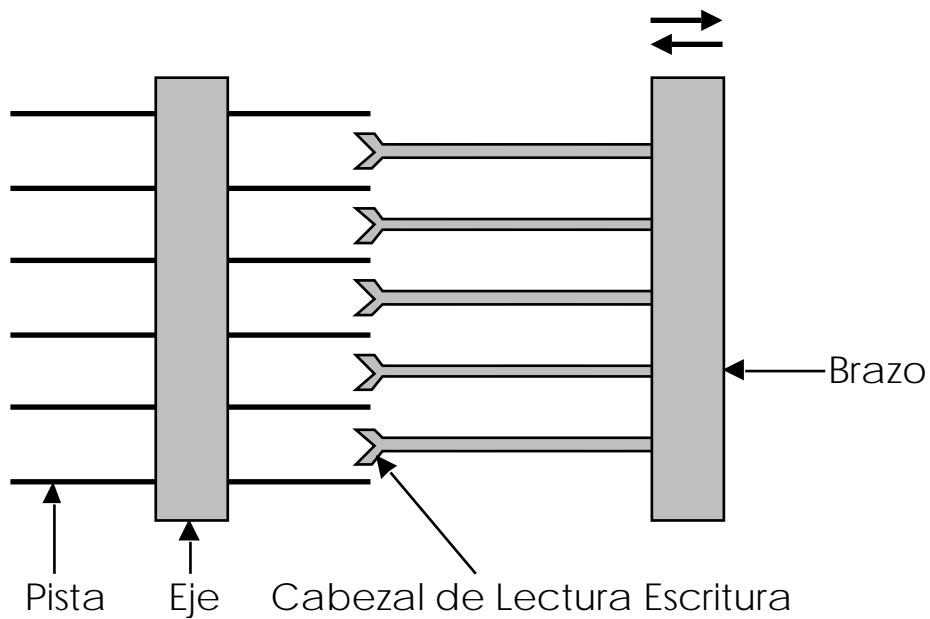
- Un disco se presenta como un conjunto de Platos, cada uno de los cuales presenta al menos una Superficie Magnética sobre la que se almacena información.
- Cada superficie se divide en Pistas, y éstas a su vez en Sectores.
- Las operaciones sobre la superficie se realiza a través del Cabezal de Lectura/Escritura.
- El movimiento del cabezal para alcanzar una pista concreta se denomina Desplazamiento.
- Cuando un disco presenta varios platos se denomina Paquete de Discos, cumpliendo una serie de propiedades,
 - La división en pistas y sectores es igual en todas las superficies, de modo que las pistas forman Cilindros.
 - Todos los cabezales se apoyan en el mismo brazo, por lo que se pueden leer datos de varias pistas dentro de un cilindro, de modo simultáneo y sin mover el brazo.
- La capacidad de los sectores suele ser constante en todo el disco, por lo que las pistas interiores presentan mayor densidad de grabación.
- Para evitar este desequilibrio, la capacidad de los sectores varía en diferentes zonas.
- El acceso a un dato en disco, lee o escribe la información de un sector sobre un buffer.

COSTE DE ACCESO A DISCOS

Organización de una Superficie



Organización de un Paquete de Discos



COSTE DE ACCESO A DISCOS

Organización por Sectores

- Un usuario asume que los ficheros aparecen en sectores contiguos dentro del disco.
- Pero esta visión no es adecuada, ya que no es posible leer sectores contiguos, porque se requiere un cierto tiempo para procesar la información inicialmente leída.
- Por tanto, si se almacenaran de este modo, sólo se podría leer un sector en cada giro.
- Para evitar este problema, se suelen intercalar los sectores lógicamente contiguos entre otros sectores, cuyo número se define por el Factor de Intercalación.
- Otra alternativa es el acceso consecutivo de un conjunto de sectores, denominado Cúmulo.
- La secuencia lógica de los cúmulos en el fichero aparece en la Tabla de Asignación de Ficheros (FAT).
- Para reducir el coste de acceso, es necesario minimizar el coste de los desplazamientos.
- Una solución es situar varios cúmulos en una zona de disco, denominada Extensión.
- De este modo, un fichero se almacena en una o varias extensiones.
- Un registro puede aparecer en un único sector, lo que puede producir Fragmentación Interna.
- En caso contrario, aumenta el coste de acceso de un registro.

COSTE DE ACCESO A DISCOS

Organización por Bloques

- Los pistas de los discos también pueden estar organizados por bloques.
- Su tamaño puede ser definido por el usuario, y su valor puede ser fijo o variable.
- Este valor se asocia al Factor de Bloque, que indica el número de registros de un fichero que se almacenarán en un bloque.
- De este modo se elimina la fragmentación antes comentada, pero quizás sea necesario más de un acceso a disco para acceder a un registro.
- En cualquier caso aparece una fragmentación a nivel de pista, ya que siempre puede quedar espacio libre al final de una pista.
- Un bloque puede descomponerse en una serie de subbloques:
 - Subbloque Contador, que incluye el número de bytes del bloque asociado.
 - Subbloque Clave, que incluye la clave del último registro del bloque.
 - Subbloque de Datos, en el que aparece la información.
- El uso del subbloque clave permite que el controlador realice búsquedas de datos.
- En cualquier caso, se puede concluir que las operaciones de acceso manejan un número de bytes definido por el usuario.

COSTE DE ACCESO A DISCOS

Overhead de las Organizaciones

- Tanto los sectores como los bloques produce cierto overhead de almacenamiento, que reduce la capacidad del disco.
- Parte de este overhead se produce cuando se formatea el disco.
- En las organizaciones por sectores, debe de incluir en cada sector la siguiente información:
 - Dirección del Sector y de la Pista.
 - Estado del Sector, útil o dañado.
 - Espacios y Marcas de Sincronización
- El usuario no tiene la posibilidad de manejar esta información.
- En las organizaciones por bloques, parte de esta información sí puede ser definida por el programador.
- Entre ésta se encuentra la proporción existente entre el tamaño relativo de los subbloque de datos y de los subbloques contador y clave.
- Esta relación se controla por el tamaño del subbloque de datos, y más concretamente por el valor del factor de bloque.

COSTE DE ACCESO A DISCOS

Cálculo del Coste de Acceso a un Disco

- El coste del acceso a un disco se calcula a partir de la suma de tres valores:
 - Tiempo de Desplazamiento, para situar el cabezal sobre el cilindro adecuado.
 - Retraso por Rotación, que incluye el tiempo necesario para situar el cabezal sobre el sector seleccionado.
 - Tiempo de Transferencia, en el que se realiza de modo efectivo la operación.
- El primero se obtiene como suma del Tiempo Inicial de Arranque del Cabezal, s , y el producto del coste de atravesar un cilindro, m , y el número de cilindros a atravesar, n .

$$f(n) = s + m \times n$$

Normalmente se utiliza un valor promedio, en el que se asume que sólo se atraviesa un tercio de los cilindros.

- La segunda depende de la velocidad de giro del disco y de la posición del cabezal, aunque se suele aproximar por el coste de medio giro.
- Por su parte, la tercera depende del número de bytes a transmitir y de la velocidad de giro.

$$\text{Tiempo Transferencia} = \frac{\text{nº bytes a transmitir}}{\text{nº bytes en la pista}} \times \text{Tiempo Rotación}$$

COSTE DE ACCESO A CINTAS

Organización de las Cintas

- Una cinta suele estar compuesta por una serie de pistas paralelas.
- Normalmente aparecen nueve pista, ocho de las cuales almacena un byte de información, y una en la que se almacena el bit de paridad del byte.
- Por tanto se puede considerar un byte como una sección de cinta de tamaño igual a un bit, y que recibe el nombre de Marco.
- Los datos se agrupan en bloques cuyo tamaño es variable, y que se asocian al valor Factor de Bloque.
- Los bloques se separan por huecos entre bloques, que son lo suficientemente grandes para posibilitar la parada y arranque de la cinta.
- Una cinta se caracteriza por los parámetros:
 - Densidad de la Cinta, medido en número de bits por pulgada, que es equivalente a bytes por pulgada.
 - Velocidad de la Cinta, medido en pulgadas por segundo.
 - Tamaño del Hueco entre Bloques.
- El factor de bloque puede tener una gran influencia en la capacidad real de la cinta, por lo que se suele utilizar también la Densidad de Grabado Efectiva.

COSTE DE ACCESO A CINTAS

Cálculo del Tamaño de una Cinta

- Para calcular el tamaño necesario de la cinta para almacenar una información se considera:
 - La Longitud Física de un Bloque de Datos.
- $$b = \frac{\text{Tamaño del Bloque (bytes por bloque)}}{\text{Densidad de la Cinta (bytes por pulgada)}}$$
- La Longitud de un Hueco entre Bloques, g.
 - El Número de Bloques, n.
 - Dados estos valores, se calcula el resultado como sigue,

$$s = n \times (b + g)$$

Cálculo del Acceso a una Cinta

- El Coste Nominal de Transmisión de Datos se define como,

$$\text{Densidad de Cinta (bpi)} \times \text{Velocidad de Cinta (ips)}$$

- Pero este valor no es demasiado realista, ya que no se tiene en cuenta el espacio ocupado por los huecos entre bloques.
- Para ello es necesario definir la densidad de grabado efectiva,

$$\frac{\text{Número Bytes por Bloque}}{\text{Número Pulgadas Necesarias para Almacenar Bloque}}$$

- Sustituyendo la densidad de cinta por este valor más realista, se obtiene el valor buscado.