CONCEPTOS FUNDAMENTALES, HISTORIA Y COMPONENTES DE UN PC

Sin duda, el PC es una de las herramientas más empleadas en la actualidad, y por ello se hace casi imprescindible conocerlo a fondo. Esta recopilación recorrerá todas las partes que componen un PC, explicando su funcionamiento, y los factores a tener en cuenta ante la compra de cada componente. En el primer capítulo se introducirán algunos conceptos fundamentales de los ordenadores. También se presentará un recorrido histórico orientado hacia el PC, y se mostrarán los componentes básicos de dicho ordenador personal. En los siguientes capítulos, se describirán cada una de las partes más importantes del PC, describiendo los elementos que los comunican.

Félix Torán Martí

Félix Torán Martí es titulado superior en ingeniería electrónica y experto en computación técnica, instrumentación virtual y desarrollo orientado a Internet. Ha desarrollado software y hardware con aplicación medioambiental. Es autor de multitud de artículos científicos y técnicos, publicados en revistas y congresos nacionales e internacionales

Recopilación de la información publicada en la revista PC WORLD

INDICE

Página

Página

Página

Página

CAPÍTULO 1.

INTRODUCCIÓN AL HARDWARE DEL PC

Es importante comenzar desde la base. La primera pregunta a plantear es la siguiente: ¿qué es un ordenador? Ante todo, es un dispositivo electrónico capaz de recibir datos de entrada, realizar una serie de operaciones con ellos, y generar datos de salida como resultado. La sucesión de cálculos a realizar está determinada por un programa.

Conviene ilustrar lo anterior mediante un ejemplo (tan sólo es preciso entender los conceptos generales, sin necesidad de introducirse en el mundo de la electrónica). Imagine por un momento que desarrolla un circuito electrónico capaz de sumar dos números (introducidos por un teclado) y presentar el resultado en una pantalla. En la Figura 1 se muestra un esquema de dicho dispositivo, cuya parte principal es un circuito encargado de recibir dos números como entrada, y devolver la suma como salida. Si ahora desea ampliar su dispositivo para que efectúe restas, deberá realizar cambios en el circuito, con toda seguridad. Aún más complicados serán los cambios si desea dotar de una funcionalidad más compleja a su dispositivo. E incluso mucho más si se debe realizar una secuencia de operaciones, una tras otra, para obtener el funcionamiento deseado.

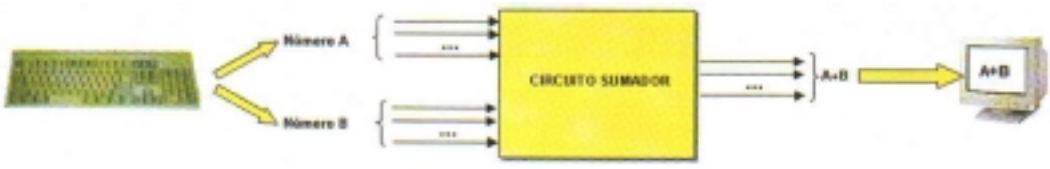


Figura 1. Un sistema electrónico sumador de números

Una solución consiste en crear un módulo electrónico capaz de realizar un conjunto de operaciones (Figura 2), que se pueden seleccionar en cada momento. Las líneas digitales de la parte inferior permiten indicar al dispositivo cuál es la operación (instrucción) a realizar en cada instante. El circuito leerá los datos de entrada, realizará la operación seleccionada, y devolverá datos de salida (empleando las líneas destinadas a ello). Una secuencia de operaciones conforma un programa, que define totalmente el comportamiento del sistema electrónico. La solución presentada ofrece una gran ventaja: es posible cambiar el funcionamiento del dispositivo simplemente modificando el programa, sin necesidad de alterar su diseño electrónico. Un ordenador responde al comportamiento presentado.

Componentes básicos de un ordenador

Una vez presentadas las ideas más básicas, es el momento de introducir los elementos fundamentales del diseño de un ordenador. Entre otros muchos,

destacan tres bloques funcionales: la memoria, la unidad central de proceso (CPU) y la entrada/salida (E/S).

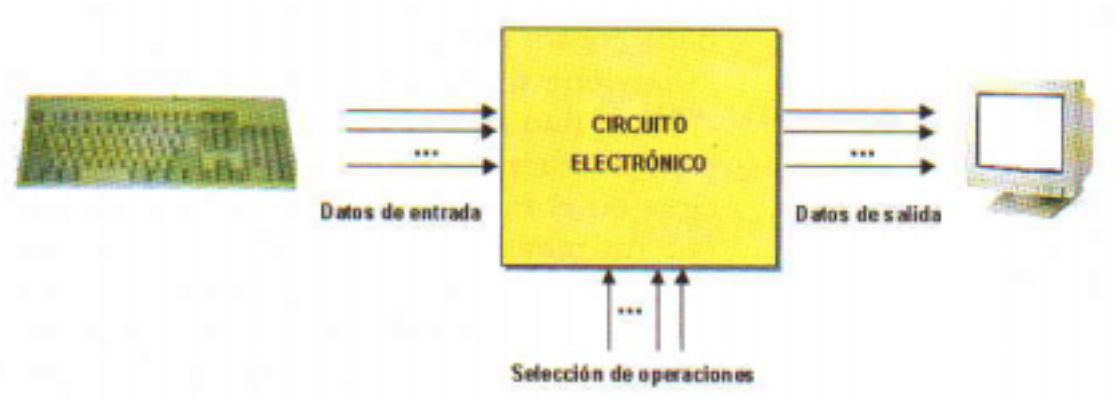


Figura 2. Sistema programable. Una primera aproximación al funcionamiento de un ordenador

La CPU es la unidad encargada de ejecutar las instrucciones definidas por los programas. La CPU comienza por leer una instrucción, la ejecuta, y genera el resultado de la operación realizada. Este proceso se repite continuamente durante su funcionamiento.

La memoria se emplea para almacenar información temporal o permanentemente. Por un lado, se almacenan las instrucciones que componen los programas. También se almacenan datos de entrada con los que debe trabajar la CPU, resultados intermedios, y datos de salida devueltos por la CPU. Por ejemplo, imagine la siguiente instrucción: "MUL A,B", capaz de multiplicar dos números A y B. La instrucción debe estar almacenada previamente en la memoria, para poder llegar a la CPU en el momento preciso. Lo mismo ocurre con A y B, que también proceden de la memoria del sistema. Tras la ejecución de la instrucción, la CPU devolverá el resultado (A x B), que se debe almacenar en la memoria para ser empleado más adelante (ya sea para enviar a un dispositivo de salida, o para utilizar como entrada en una futura instrucción).

La E/S es el medio establecido para la comunicación de la CPU con su entorno exterior (es más un concepto que algo palpable). Mediante la definición de una cierta interfaz, la CPU puede recibir datos de los dispositivos que la rodean (denominados periféricos), operar con dicha información, y enviar los resultados generados hacia dichos dispositivos.

Como habrá intuido, los componentes de un ordenador deben hallarse conectados entre sí. El elemento encargado de dicha tarea es el bus. Un bus se debe entender como un grupo de "cables" (líneas digitales) que interconectan los bloques funcionales de un ordenador, permitiendo la interacción entre los mismos. Visto de otro modo, los componentes se enlazan al bus para conectarse así con el resto de los elementos. Ya que el bus une a todos los elementos entre sí, podrían aparecer conflictos si varios de ellos intentan utilizar el bus al mismo tiempo. Esto obliga a establecer una regla importante: en cualquier instante, sólo un componente puede colocar información en el bus. La Figura 3 resume todo lo explicado, mostrando el esquema básico de un ordenador.

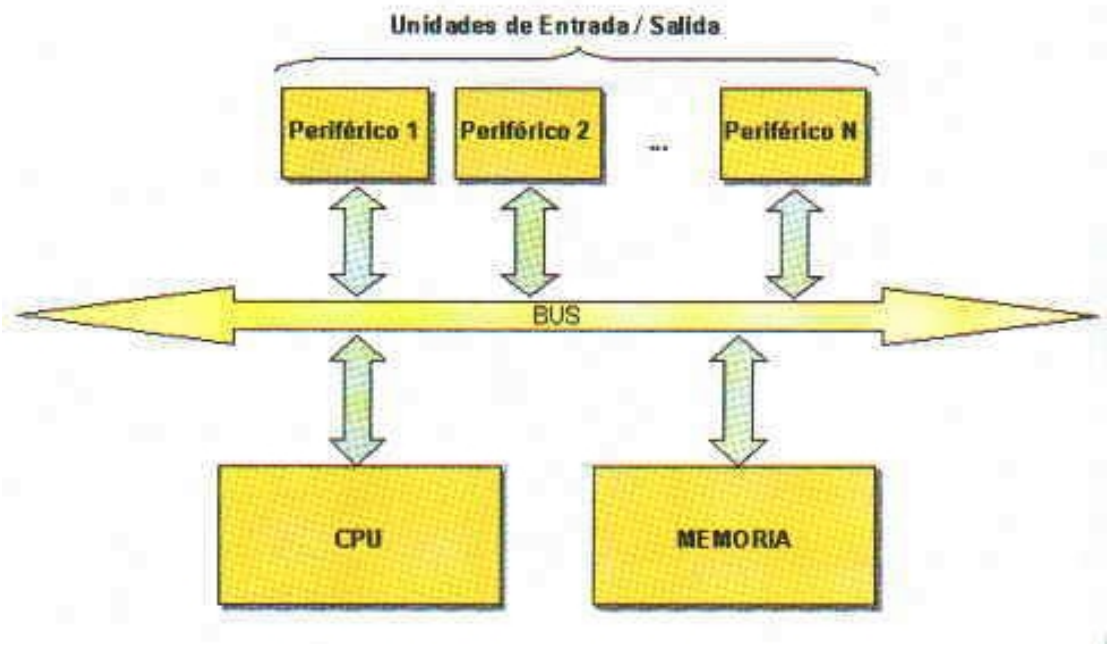


Figura 3. Esquema básico de un ordenador

Arquitecturas básicas

Los ordenadores se pueden clasificar basándose en su arquitectura. Existen dos arquitecturas principales, denominadas Von Neumann y Hardvard. La diferencia fundamental se encuentra en el modo de almacenar en la memoria las instrucciones y los datos con los que trabajan.

En las máquinas de Von Neumann, las instrucciones y los datos conviven en el mismo espacio de memoria, sin existir separación física.

Los ordenadores con arquitectura Hardvard dividen el espacio de almacenamiento en dos bloques de memoria físicamente separados. Uno de los bloques almacena las instrucciones, y el otro almacena los datos. El acceso a dichos espacios de almacenamiento se realiza mediante buses diferentes, lo que hace posible la lectura simultánea de instrucciones y datos.

Las máquinas con arquitectura Hardvard presentan un mayor rendimiento en la ejecución de instrucciones, ya que pueden leer instrucciones y datos de forma simultánea. Hay que tener presente que en una memoria sólo se puede obtener un dato o instrucción en cada acceso (salvo en el caso de las memorias multipuerto, que se abordarán en próximas entregas). Para aclarar esto, considere el proceso de ejecución de una instrucción en una máquina Von Neumann:

Primero se accede a la memoria para obtener la instrucción a ejecutar, y se descodifica dicha instrucción, conociendo así la operación a realizar y los operandos con los que trabajar.

Después se realizan los accesos necesarios a la memoria (uno tras otro) para obtener los datos con los que operar. Por ejemplo, en la instrucción de ejemplo "MUL A,B", se requieren dos accesos a la memoria: uno para obtener el valor A, y otro para obtener el valor B. Finalmente, se ejecuta la instrucción, y se accede a la memoria para almacenar el resultado de la operación.

En una máquina con arquitectura Hardvard, mientras la CPU obtiene los datos requeridos por una instrucción se puede leer -simultáneamente- la siguiente instrucción a ejecutar, con lo que el rendimiento es claramente superior. Evidentemente, esto no se puede hacer en una máquina Von Neumann, ya que mientras se accede a los datos, no es posible leer una instrucción al mismo tiempo, puesto que sólo hay una memoria, y un solo bus que la une a la CPU.

También existen máquinas con arquitectura Hardvard modificada, que emplean dos buses diferentes para acceder a los datos. El funcionamiento mejora aun más, puesto que se puede leer una instrucción y dos datos de forma simultánea.

Es importante recordar que el PC es una máquina Von Neumann, y por tanto cumple con las características descritas para dicha arquitectura.

Microprocesadores.

Aunque ya se estudiará en posteriores capítulos, es conveniente comentar las partes que componen la CPU:

Unidad de descodificación de instrucciones. Se encarga de interpretar las instrucciones que componen los programas. A partir de una instrucción, extrae la acción a realizar y la forma de encontrar los datos con los que trabajar (almacenados en memoria, incluidos en la propia instrucción, etc.).

Unidad Aritmético Lógica (más conocida por el acrónimo inglés ALU). Es el módulo encargado de efectuar las operaciones aritméticas (suma, resta, etc.) y lógicas (and, not, or, etc.).

Registros. Cada registro es una unidad de memoria, que permite almacenar temporalmente un dato. La CPU puede leer los registros para obtener datos de entrada. También puede almacenar los resultados devueltos, de forma que futuras instrucciones puedan emplearlos como entrada. En todo ordenador existen algunos registros especiales, dedicados a tareas muy concretas (contador de programa, indicadores del estado de la CPU, etc.).

Unidad de control de buses. Este bloque controla los procesos de transferencias de información, ya sea internamente a la CPU, o de forma externa.

Antiguamente, estos bloques funcionales se implementaban mediante complicados circuitos electrónicos, que ocupaban un espacio considerable y ofrecían un rendimiento no muy elevado (en comparación con el actual). Con el avance de la tecnología, nacieron los microprocesadores, circuitos integrados que contienen una CPU completa. Esta invención marcó por completo la historia de los ordenadores, dando paso al nacimiento de los ordenadores personales

Breve historia de los ordenadores personales

Los ordenadores personales nacieron con el objetivo de acercar los ordenadores a pequeñas empresas y a los hogares, donde antes era imposible

disponer de uno. Tras la invención del microprocesador, dicha intención estaba cada vez más cerca de hacerse realidad.

Un primer acontecimiento importante ocurría en 1971: Intel desarrollaba su primer microprocesador, el 4004. Se trataba de un procesador de 4 bits, que reducía considerablemente la necesidad de implementar circuitos adicionales para su funcionamiento. La misma empresa anunció el lanzamiento de un chip de memoria RAM de 1 kbit, cantidad superior a la usual en aquella época.

En 1972, Intel lanzó los procesadores 8008 y 8080, en este caso de 8 bits y capaces de acceder a 16 kB de memoria.

Figura 4. Oredenador Altair 8800

Un desarrollo clave se realizó en el centro de investigación de Xerox en Palo Alto (PARC), en el año 1973. Un equipo desarrolló un ordenador con un aspecto que se aproximaba bastante al de un PC actual, y al que se bautizó con el nombre de Alto. Incluso disponía de un dispositivo de entrada muy similar a un ratón. Lamentablemente, el estado de la tecnología en aquel momento impidió el éxito del proyecto.

El microprocesador 8080 dio lugar a diversos desarrollos, como es el caso del ordenador Altair 8800 (Figura 4), que alcanzó un gran éxito. Esta máquina carecía de teclado y pantalla (tan sólo presentaba interruptores y luces), pero su arquitectura basada en la conexión de tarjetas dio lugar al bus estándar S-100 más tarde. En 1975 ya se había presentado una implementación del lenguaje BASIC para el ordenador Altair (hecho en el que Bill Gates tuvo una importante participación). Otros ordenadores que se diseñaron a partir del procesador 8080 fueron IMSAI 8080 y Mark-8.

En 1976, Intel anunciaba su nuevo procesador 8085, capaz de trabajar a 5 MHz y con un bus de 8 bits. Por su parte, Zilog introducía el procesador Z80, basado en el 8080 de Intel. Ese mismo año, MOS Technology introduce el procesador 6052. Basado en este procesador, se lanza el Apple-I, que fue después sucedido por el Apple-II.

Como se puede observar, iban apareciendo nuevos microprocesadores, y al tiempo se iban desarrollando nuevos ordenadores personales basados en dichas CPU.

En la década de los 80 imperaban ordenadores personales como ZX-Spectrum (sucesor del ZX-80 y ZX81, creados por Sir Clive Sinclair). Commodore lanzó su VIC-20, y posteriormente el Commodore 64 (a partir del procesador 6052). Estos ordenadores se basaban en procesadores de 8 bits y manejaban un máximo de 64 kB de memoria. Amstrad (con su CPC) y los ordenadores MSX intentaron rebasar ese límite, manteniendo los chips empleados, pero no lograron el éxito deseado.

El nacimiento del PC

El PC fue introducido en 1981 por IBM. Se presentaba como un ordenador pesonal basado en el microprocesador Intel 8088, de 16 bits. Este primer PC (Figura 5) constaba de 16 kB de memoria, ampliables a 64 ó 256 kB: El almacenamiento externo se realizaba mediante cintas de cassette, que después se transformaron en unidades de disco de 5,25 pulgadas.

En 1983 se lanza el IBM PC-XT, cuya principal novedad era un disco duro de 10 MB. En 1984 aparece el IBM PC-AT, basado en el nuevo procesador 286. Ofrecía ranuras de expansión de 16 bits (empleando el estándar industrial ISA) y 20 MB de disco duro. Todos los sucesores de dicho ordenador quedan dentro de la categoría AT, aunque actualmente se hace referencia a ellos bajo el nombre de PC.

Figura 5. Ordenador IBM PC

El IBM PC-AT experimenta un gran éxito en los siguientes años. En 1986, IBM lanza la siguiente generación de ordenadores PC, esta vez basados en el procesador 386. Con aquel procesador, la competencia comenzó a adelantarse a IBM, creando ordenadores compatibles que ofrecían características más ventajosas, y que eran incompatibles con las ofrecidas por IBM. El PC comienza a desligarse de IBM. De hecho el primer 386 del mercado no fue presentado por IBM, sino por Compaq.

A continuación, el PC sigue su evolución, alcanzando un éxito espectacular: está presente en infinidad de hogares y empresas de todo tipo, y mueve un mercado de software sin precedentes. Cada nueva generación del PC viene definida por una nueva versión de su procesador Intel: 486, Pentium, Pentium II, Pentium III y el nuevo Pentium N.

En la actualidad, no cabe duda de que el PC sigue siendo una estrella en el mundo de los ordenadores personales, y su futuro se presenta inmejorable.

Componentes del PC

Los elementos básicos de un ordenador ya se han mostrado en la Figura 3. En la realidad, los ordenadores presentan algunos componentes adicionales, y el PC no podía ser menos. La Figura 6 muestra una visión global de los elementos que forman un PC.

El primer componente a destacar es un circuito impreso, al que comúnmente se denomina "placa base". La placa base es el lugar donde van conectados todos los elementos del PC. Sobre dicha placa se pueden encontrar los elementos básicos comentados anteriormente: CPU, memoria y buses del sistema. Además, existen otros componentes, que se citan a continuación:

Circuitos controladores del sistema. Conjunto de chips que se encargan de controlar el tráfico de información en el interior del PC.

BIOS. Constituye la interfaz entre el hardware y el software del sistema. Es el módulo encargado de cargar el sistema operativo al arrancar el PC, y además permite configurar los parámetros de funcionamiento de la máquina. La BIOS contiene el programa de menor nivel de abstracción que se ejecuta en un PC.

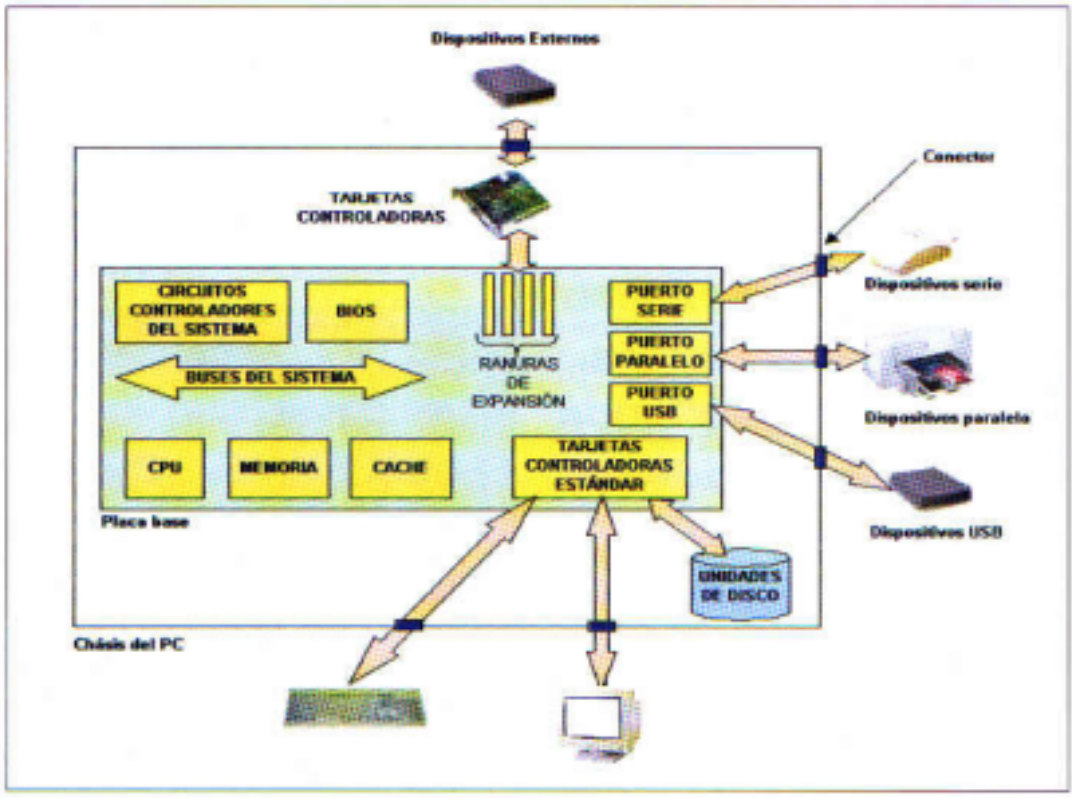


Figura 6. Los componentes de un PC

Memoria caché. Es un tipo de memoria con poco espacio de almacenamiento, pero de acceso muy rápido. En la caché se almacena la información que con gran probabilidad va a necesitar acceder la CPU, de forma que se mejora el rendimiento del sistema.

Puertos serie, paralelo, USB, etc. Constituyen un canal de comunicación con dispositivos externos, como el ratón, la impresora, etc.

Controladores estándares. Son dispositivos encargados de la comunicación con los dispositivos estándares, como el teclado, el monitor y las unidades de disco.

Ranuras de expansión. Permiten instalar en el sistema nuevos dispositivos de E/S adicionales, ampliando así las posibilidades del PC: unidades de almacenamiento externas, tarjetas de sonido, módems internos, dispositivos de captura de imágenes, tarjetas de red, etc. Normalmente, en las ranuras se inserta una tarjeta controladora, que proporciona uno o varios conectores para el enlace con un dispositivo externo.

Saliendo un poco más hacia el exterior, dentro del chasis del ordenador se encuentran dispositivos como las unidades de disco, y algunos subsistemas auxiliares como la fuente de alimentación, ventiladores, etc.

Ya en el exterior del sistema, se encuentran los elementos estándares con los que interacciona el usuario: teclado, monitor, impresoras, etc. Dichos elementos se unen a la placa base a través de tarjetas controladoras, o bien empleando los puertos serie, paralelo o USB.

CAPÍTULO 2.

LA UNIDAD CENTRAL DE PROCESO (I)

Sin ninguna duda, el elemento estrella de un ordenador personal es la unidad central de proceso (CPU). Tanto es así, que la primera característica que se tiende a emplear para comparar ordenadores, es el modelo de CPU y la frecuencia de reloj (MHz) a la que es capaz de trabajar. Sin embargo, si se acude a las fuentes de información habituales (comparativas, información proporcionada por los fabricantes, etc.), se aprecia la existencia de multitud de términos profundamente técnicos. Dichos conceptos son los que realmente permiten comparar procesadores, o determinar si un procesador es adecuado para una aplicación específica, y requieren de un conocimiento básico acerca de lo que ocurre en el interior de la CPU. Por ello, este capítulo inicia un recorrido a través del interior de los procesadores, mostrando los componentes fundamentales, su funcionamiento y los conceptos más importantes a tener en cuenta.

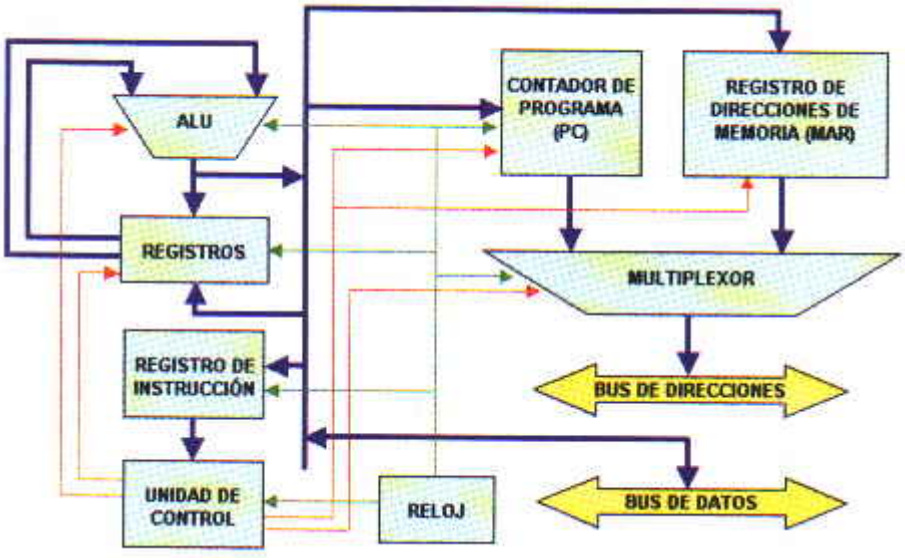


Figura 1. Estructura de un procesador sencillo. Se pueden apreciar los componentes básicos de cualquier procesador

Estructura básica de la CPU

La Figura 1 muestra la estructura interna de un procesador sencillo con arquitectura Von Neumann. El diseño es básico (se corresponde con el presentado por algunos procesadores sencillos de 8 bits) pero su comprensión permitirá entender el funcionamiento de procesadores más complejos, que amplían la estructura presentada.

Se pueden apreciar los siguientes elementos:

Bus de direcciones. Permite enviar direcciones a la memoria y periféricos. Las direcciones son números que indican la posición de memoria donde se desea escribir un dato, o de donde se desea leerlo. La memoria se abordará con detalle en próximos capítulos, al igual que los periféricos.

Bus de datos. Permite al procesador enviar datos a la memoria y periféricos. De la misma forma, permite que el procesador reciba datos de dichos elementos. Para escribir (o leer) un dato, primero es necesario colocar la dirección donde escribirlo (o de donde leerlo) en el bus de direcciones.

Registro de instrucción. En cada momento, este registro almacena la instrucción que está siendo ejecutada por la CPU.

Archivo de registros. Son almacenes temporales de datos, de acceso muy rápido. En general, almacenan los operandos sobre los que actúan las instrucciones, y también los resultados de éstas (para un futuro uso). También existen registros de aplicación específica, como se verá más adelante.

Unidad Aritmético-Lógica. Más conocida por el acrónimo inglés ALU (*Arithmetic-Logic Unit*). La ALU es el motor de cálculo del procesador, ya que se encarga de realizar las operaciones para las que está capacitado. Como se aprecia en la Figura 1, la ALU es capaz de tomar 2 datos como operandos y producir una salida, resultado de aplicar una operación (seleccionable mediante unas líneas de control). Las operaciones soportadas pueden ser aritméticas (suma, resta, etc.), lógicas (and, or, etc.) y otras operaciones como desplazamientos de bits. Sin duda, la ALU es un elemento muy importante, ya que define las operaciones de cálculo que la CPU puede realizar.

Contador de programa. Se trata de un registro especial, al que se denomina normalmente PC (del inglés *Program Counter*). El PC contiene, en cada momento, la dirección de memoria en la que se encuentra la siguiente instrucción a ejecutar. Cada vez que se obtiene una nueva instrucción de la memoria, el PC se actualiza para apuntar a la próxima instrucción a ejecutar. Normalmente, dicha actualización consiste en un simple incremento, para apuntar a la celda de memoria siguiente (téngase en cuenta que, en general, las instrucciones de un programa se almacenan en posiciones de memoria consecutivas). Este comportamiento cambia cuando se ejecuta una instrucción de salto. En ese caso, el PC se actualiza con la dirección de destino del salto, donde se encuentra la siguiente instrucción a ejecutar.

Registro de direcciones de memoria. También denominado MAR (*Memory Address Register*). Este registro almacena la próxima dirección de memoria de la que leer (o en la que escribir) un dato. Se trata de un registro que complementa al PC: mientras que este último apunta a instrucciones, el MAR apunta a datos.

Multiplexor. No se trata de un elemento exclusivo de las CPU, sino de una pieza fundamental del diseño de sistemas digitales. En el esquema de la Figura 1, el multiplexor actúa como una especie de conmutador, transporta el contenido del PC o del MAR hacia el bus de direcciones. La línea roja de la parte izquierda es el terminal que permite seleccionar si es el contenido del PC o el del MAR el que llegará al bus (colocando en él un 1 o un 0 lógico). En la Figura 2 se ilustra este hecho.

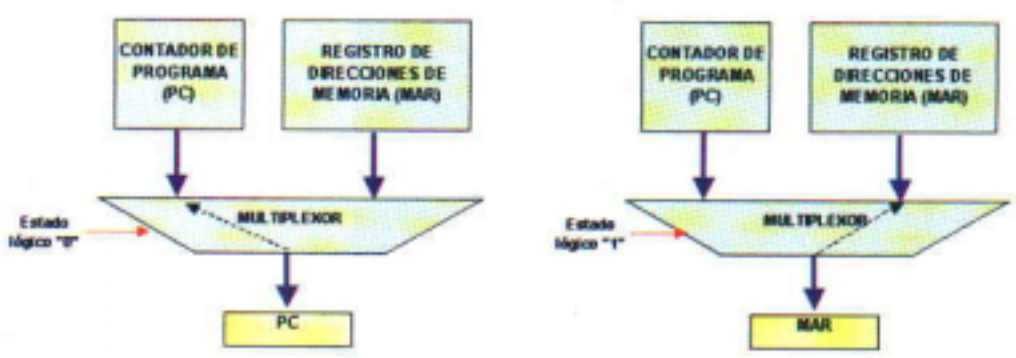


Figura 2. Funcionamiento del multiplexor presente en la figura 1

Reloj. En la actualidad, prácticamente todos los procesadores son sistemas digitales síncronos. Esto significa que trabajan al ritmo marcado por una señal de reloj. Dicha señal no es más que un conjunto de pulsos distanciados por igual en el tiempo. Más adelante se comprobará que una mayor velocidad de reloj no siempre implica un mayor rendimiento del procesador.

Unidad de control. Es el bloque encargado de coordinar el funcionamiento interno de la CPU. Indica a cada componente interno cómo debe funcionar y cuándo tiene permiso para entrar en acción. En otras palabras, la unidad de control se puede entender como el director de orquesta de la CPU.

Tras los anteriores comentarios, se ha obtenido una visión de sistema de la CPU. Como en todo sistema, se dispone de bloques que cooperan entre sí para lograr una labor común (en este caso, ejecutar programas). En el siguiente apartado se muestra en qué consiste dicha cooperación.

Funcionamiento

Considere una posible instrucción a ejecutar por el procesador de la Figura 1 (denominada, por ejemplo, "ADD Rl, R2, R3"). Suponga que dicha instrucción toma los datos almacenados en los registros denominados R 1 y R2, realiza la suma, y almacena el resultado en el registro R3. Como ya habrá intuido, R 1, R2 y R3 se encuentran en el archivo de registros. El proceso completo para ejecutar dicha instrucción es el siguiente:

1. El PC se actualiza con la dirección de memoria donde se halla la siguiente instrucción a ejecutar ("ADD R1, R2, R3").
2. La unidad de control (UC) modifica el estado del multiplexor, de forma que el contenido del PC se dirija hacia el bus de direcciones.
3. La memoria recibe la dirección contenida en el PC, y responde devolviendo el dato contenido en dicha celda a través del bus de datos. Dicho dato representa a la instrucción "ADD R1, R2, R3" mediante un valor numérico binario.
4. La instrucción recibida se almacena en el registro de instrucción.
5. La UC descodifica la instrucción. Es decir, averigua cuál es la instrucción (ADD) y cuáles son los operandos sobre los que ésta trabaja (Rl, R2 y R3). Se debe tener presente que una instrucción se representa por un número

binario, que ofrece información sobre la operación a realizar (mediante un un grupo de bits), y los operandos sobre los que trabajar (en otros grupos de bits). Para facilitar su comprensión, las instrucciones se suelen etiquetar mediante mnemónicos (por ejemplo, "ADD"), formando lo que se denomina "lenguaje ensamblador", propio de cada procesador.

6. La UC se comunica con el archivo de registros (véase la conexión en rojo), y hace que el contenido de los registros R1 y R2 vaya a parar a las 2 entradas de la ALU.
7. La UC configura la ALU (mediante la línea de control en rojo, que realmente son varias líneas) indicando que se desea realizar la operación suma.
8. La UC configura el archivo de registros, de forma que el resultado de la suma se conduzca hacia el registro R3.
9. La UC ordena a la ALU que realice la operación actualmente configurada. Tal y como se pretendía, en el registro R3 se dispone de la suma de los valores contenidos en Rl y R2, con lo que la instrucción ha sido completada.

Como conclusión importante del ejemplo, conviene remarcar que la ejecución de una instrucción sencilla se compone de una serie de operaciones internas a la CPU, todas controladas por la UC. Esta última sabe cuál es la secuencia de tareas a aplicar para cada instrucción, tras la etapa de descodificación. Como imaginará, las operaciones a realizar para ejecutar una misma instrucción dependen de la arquitectura de la CPU en uso. Imagine dos CPU con diferente diseño, pero que soportan el mismo juego de instrucciones. En ese caso, un mismo programa sería válido para ambas, y daría lugar a los mismos resultados. Pero es importante notar que el rendimiento del programa podría ser diferente en cada CPU, ya que el hardware es diferente (por ejemplo, el número de tareas internas a realizar y la duración de las mismas puede ser diferente). En otras palabras: el juego de instrucciones condiciona la compatibilidad de los programas, pero el diseño de la CPU decide el rendimiento.

En este punto, ya se han introducido los componentes fundamentales de la CPU. Los siguientes apartados muestran con mayor detalle dos elementos destacados: la ALU y los registros.

La Unidad Aritmético-Lógica

La ALU constituye el motor de cálculo de la CPU. El conjunto de operaciones que la ALU es capaz de efectuar, y la rapidez con la que puede realizarlas, influyen en gran medida en el rendimiento final del microprocesador. Por ello, el diseño de la ALU es crucial en el diseño de un procesador y merece una atención especial.

La Figura 3 muestra el esquema general de una ALU. Los terminales **A** y **B** son en realidad puertos de entrada (compuestos por varias líneas, formando sendas palabras binarias). Los datos introducidos en **A** y **B** son los operandos con los que la ALU realizará la siguiente operación. Dicha operación se selecciona mediante las líneas del puerto denominado OPERACIÓN. Cada posible operación a realizar tiene asociado un número binario, que es el que se introduce en dicho puerto. Cuando la ALU realiza la operación matemática o lógica seleccio-

nada, devuelve el resultado empleando el puerto **C**. El resto de líneas de salida son bits (comúnmente denominados banderas o *flags*) que indican el estado de la última operación. En la Figura 3 se han indicado los *flags* **C** (se ha producido acarreo), **N** (resultado negativo), **Z** (el resultado es cero) y **V** (se ha producido un desbordamiento, es decir, se ha superado el valor más alto que es posible representar).

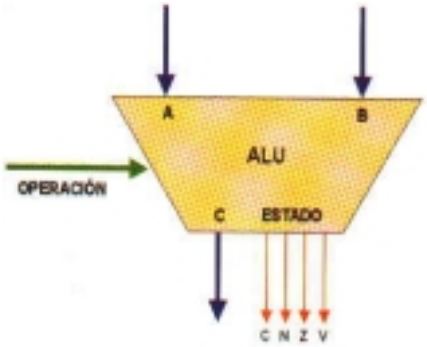


Figura 3. Estructura general de una Unidad Aritmético-Lógica (ALU)

En el diseño de una CPU, es muy importante considerar la complejidad que se va a asignar a la ALU. Algunos procesadores contienen una ALU que soporta operaciones como la multiplicación y división, e incluso funciones matemáticas como el seno, lo que consigue acelerar notablemente la ejecución de los programas. Como era de esperar, a mayor complejidad, mayores dimensiones presenta la ALU, dejando menos espacio útil para otras partes de la CPU. Realmente, si en la aplicación final las operaciones van a ser sencillas, conviene implementar una ALU también sencilla, que dejará más espacio para otros bloques importantes de la CPU. Por ejemplo, se puede prescindir de la multiplicación, realizándola por software como una sucesión de sumas (lo que resulta lento, pero si no es frecuente, no tiene un gran impacto en el rendimiento general).

Los registros

Como ya se ha introducido, los registros son celdas capaces de almacenar un dato temporalmente. Pero, ¿cuál es el motivo de su existencia? Si la CPU tuviera que trabajar directamente con los datos almacenados en la memoria principal, el funcionamiento se haría más lento de lo necesario (son necesarios varios ciclos de reloj para recuperar o escribir un dato). Hay que tener en cuenta que la memoria es un elemento externo, y, por tanto, requiere un tiempo de acceso que resulta elevado. En cambio, el acceso a los registros es muy rápido, ya que se encuentran en el interior del procesador y su diseño los hace rápidos (se accede en un ciclo de reloj). Por ello, el procedimiento general consiste en escribir los datos (bien desde la memoria o de forma directa) en los registros, operar con ellos, y almacenar los resultados también en los registros. Cuando se obtiene un resultado que se desea almacenar, se copia desde los registros hasta la memoria principal. El archivo de registros es una memoria de tipo multipuerto, ya que puede direccionar para que devuelva el contenido de dos celdas de memoria simultáneamente. Gracias a ello -como se ha apreciado

antes- la ALU puede acceder simultáneamente al contenido de dos registros. La Figura 4 muestra la estructura de un archivo de registros.

Además de los registros de propósito general, existen algunos con misiones específicas. Uno de los más conocidos es el denominado acumulador. Se encarga de sumar valores a su contenido, y actualizarse con el resultado. Tal es su importancia, que existen procesadores muy simples que son capaces de funcionar con este registro únicamente. Otro registro especial es el contador de programa, que se ha comentado anteriormente. También existen registros que se utilizan como indicadores, donde cada bit de su contenido es un *flag* que indica el estado de un aspecto concreto de la máquina. Por ejemplo, se suele indicar la existencia de acarreos, resultados nulos, desbordamientos, etc. También es usual disponer de registros tipo pila, usualmente LIFO (*Last In - First Out*, el último en entrar es el primero en salir). Al introducir datos, estos se "apilan" uno sobre el otro, y se recuperan en orden inverso al empleado al introducirlos. La Figura 5 muestra el funcionamiento de dicho tipo de registros. Su utilidad es vital en las llamadas a subrutinas: fragmentos de programa a los que se salta, son ejecutados y, tras ello, se continúa la ejecución en el punto donde se produjo el salto. Cada vez que se produce un salto a una subrutina, se introduce el valor del PC en la pila (esto es, la dirección a la que retornar). Al finalizar la ejecución de la subrutina, se extrae de la pila la dirección a la que retornar y se coloca en el registro PC. Si dentro de una subrutina se llama a otra, no hay problema, ya que las direcciones de retorno se sacarán de la pila en orden inverso al que entraron.

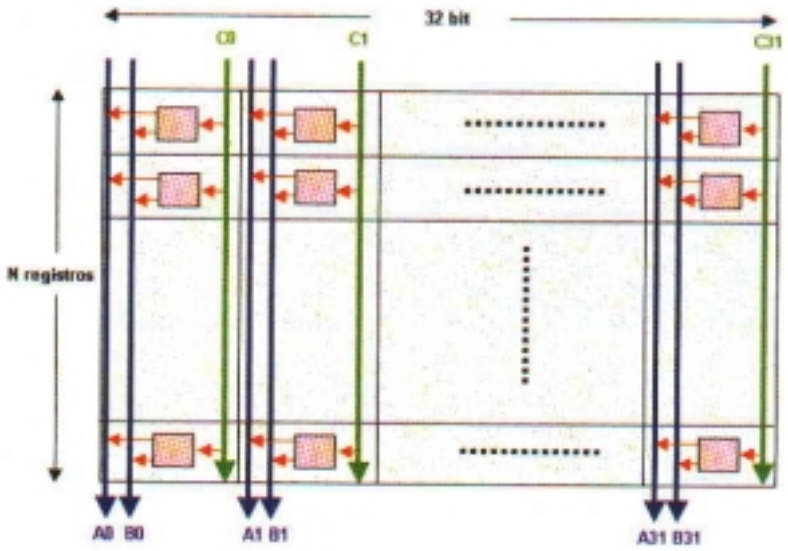


Figura 4. Estructura típica de un archivo de registros. Básicamente, es una matriz de celdas de memoria capaces de almacenar un bit (valor 0 o 1). Cada fila de la matriz forma un registro, y puede almacenar 32 bits (columnas). Cada dato a almacenar se introduce mediante los terminales C0, C1, …, C31. Los terminales A0, …, A31 y B0, …, B31 permiten leer el contenido de dos registros simultáneamente (de acuerdo con las necesidades de la ALU)

Arquitecturas RISC y CISC

Basándose en el juego de instrucciones soportado, los procesadores se clasifican en dos grandes grupos:

Procesadores RISC *(Reduced Instruction Set Computers).* Este tipo de procesadores presenta un juego de instrucciones reducido, y en el que cada instrucción realiza una tarea sencilla. La ejecución de una instrucción de alto nivel de complejidad se puede llevar a cabo a partir de una secuencia de las sencillas instrucciones disponibles.

Procesadores CISC *(Complex Instruction Set Computers).* Ofrecen un juego de instrucciones extenso y complejo. En muchos de ellos, cada instrucción compleja se ejecuta internamente mediante un microprograma, compuesto por instrucciones que operan sobre los elementos internos del procesador. Por ejemplo, operaciones como "decrementar un registro y efectuar un salto si el resultado es cero", se encuentran englobadas bajo una única instrucción en los procesadores CISC. Los procesadores Intel Pentium y Motorola 680X0 son ejemplos de procesador CISC.

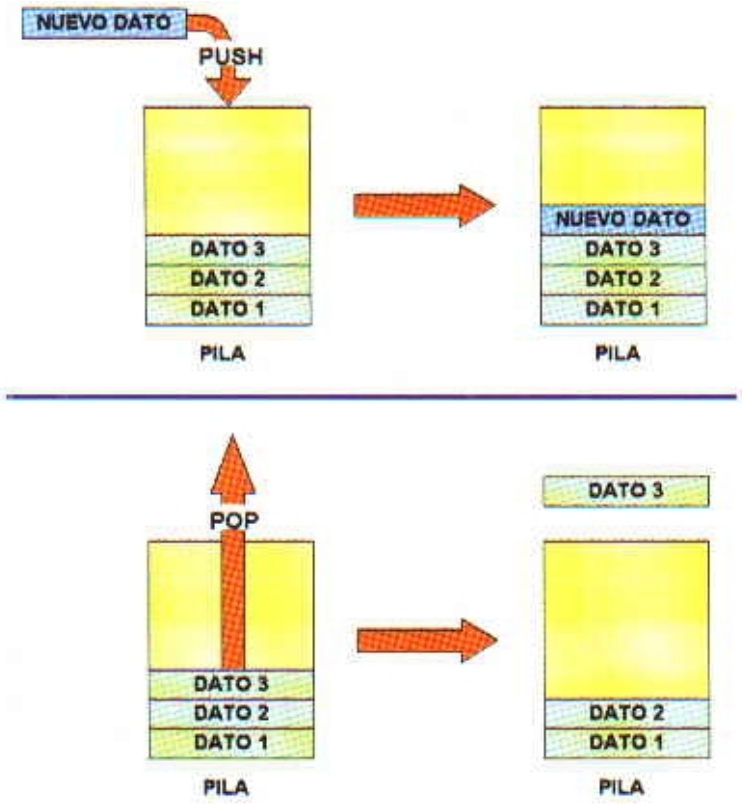


Figura 5. Funcionamiento de un pila FIFO. En la parte superior se ilustra el proceso de introducción de un nuevo dato en la pila (operación denominada PUSH). En la parte inferior se muestra el proceso de extracción de un dato de la pila (operación POP)

En primer lugar, se debe tener en cuenta que cada nueva instrucción añadida a una CPU implica añadir circuitería adicional, y, por tanto, se requiere de mayor espacio a ocupar en el chip. Una circuitería más compleja hace que el procesador trabaje con mayor lentitud. Por supuesto, el aumento de la complejidad de las instrucciones empeora aún más la velocidad de ejecución (ya que cada tarea se compone de multitud de tareas de bajo nivel) y aumenta la circuitería necesaria.

Por tanto, soportar una gran cantidad de complejas instrucciones se paga con un procesador mayor y más lento. Todo esto parece, pues, indicar que es preferible optar por los procesadores RISC, ya que al soportar menor número de instrucciones y ser éstas más simples, se conseguirá un mayor rendimiento con menor espacio. Si se trabaja con un lenguaje de alto nivel (por ejemplo, C++), el compilador se encargará de traducir las complejas instrucciones de dicho lenguaje, en una secuencia de sencillas instrucciones soportadas por el procesador, que se ejecutarán de forma eficiente.

Entonces, ¿por qué existen los procesadores CISC? Estas CPU tenían un gran sentido en los años 70 y 80, cuando las memorias eran lentas y ofrecían poco espacio de almacenamiento. Con una máquina CISC, era posible emplear instrucciones que englobaban multitud de tareas, con lo que los programas se almacenaban en un reducido espacio de memoria. Actualmente, ambos conceptos se hallan cada vez más solapados. La realidad es que muchos de los procesadores RISC actuales soportan tantas instrucciones como los antiguos procesadores CISC. Por otra parte, los procesadores CISC actuales usan multitud de técnicas, que más bien pertenecen a los procesadores RISC.

CAPÍTULO 3

LA UNIDAD CENTRAL DE PROCESO (II)

Al leer cualquier comparativa actual entre procesadores, los términos "pipelining" o "pipeline" siempre están presentes. No es de extrañar, ya que se trata de una técnica para mejorar el rendimiento de una CPU. Su implementación supone un enfrentamiento con multitud de problemas que es importante conocer. El conocimiento de esta técnica permite comprender mejor cuáles son los factores que definen la potencia real de una CPU, y aporta un arma importante a la hora de evaluar procesadores. Como apreciará en este capítulo, la comparación de procesadores basándose en su velocidad de reloj está lejos de ser realista, en la mayoría de los casos. Los siguientes apartados muestran con detalle esta interesante técnica, partiendo desde la base, y proporcionando información que cualquier comprador de un procesador debería tener en cuenta.

Introducción al pipelining

Como ya se apreció en el anterior capítulo, la CPU realiza una sucesión de tareas para ejecutar cada instrucción. En principio, se puede entender que la CPU realiza toda la secuencia de tareas para cada instrucción, antes de pasar a repetir el proceso completo con la instrucción siguiente. Pero, ¿no sería posible la aplicación de cierto paralelismo (funcionamiento simultáneo de algunas o todas las etapas del sistema), de forma que aumente el rendimiento? En efecto, esto se aplica en las CPU mediante la técnica denominada pipelining.

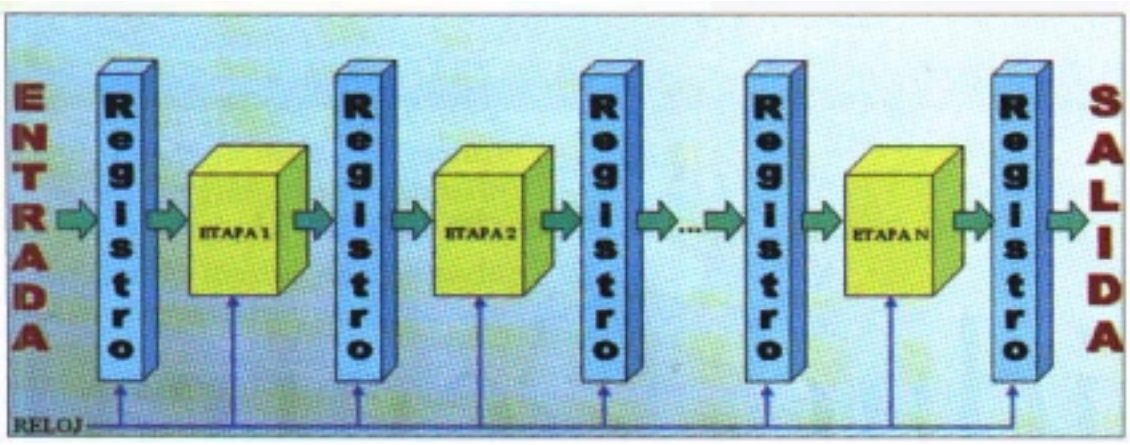


Figura 1. Estructura de una pipeline

El primer paso que aplica dicha técnica es justamente lo que se ha introducido: desglosar el funcionamiento de la CPU en diversas etapas, a las que se asignan unidades de procesamiento independientes. Dichas etapas se ejecutan de forma secuencial, y siguiendo un orden fijo. De esta forma, se obtiene lo que se denomina "pipeline" (tubería). En el fondo, este planteamiento no deja de ser

una aplicación del famoso principio "divide y vencerás", y por tanto sugiere que el rendimiento de la CPU se puede mejorar de esta forma.

Tras aplicar la última de las etapas, la instrucción en curso habrá sido ejecutada por completo. El paso de una etapa a la siguiente lo controla la señal de reloj de la CPU. La Figura 1 ilustra la estructura de una pipeline. Se aprecia la existencia de registros entre las distintas etapas. A través de estos, cada etapa puede enviar información a la siguiente.

Un símil interesante es una cadena de ensamblaje de vehículos, donde cada etapa realiza una tarea diferente (colocar ruedas, fijar el chasis, aplicar pintura, etc.), obteniendo un vehículo completo tras la última etapa.

Continuando con la cadena de ensamblaje, imagine que un vehículo no entra en la cadena hasta que el vehículo anterior ha abandonado la última etapa. En ese caso, el rendimiento no será óptimo en absoluto (piense que los operarios de cada etapa tendrán un generoso periodo de descanso hasta recibir el próximo vehículo, y por tanto no están trabajando todo lo que sería deseable). Este comportamiento no es más que el presentado hasta el momento, y es propio de los procesadores que no emplean pipelines.

El pipelining se alcanza aportando un ingrediente más a la cadena de ensamblaje: se introducirá un nuevo vehículo en la primera etapa cada vez que ésta quede libre. Lo mismo se hará en el resto de las etapas: de forma simultánea, cada vehículo pasa de una etapa a la posterior. Por lo tanto, tras un llenado inicial, la cadena siempre contiene vehículos en todas las etapas, y se produce un desplazamiento continuo desde la entrada hacia la salida. Cada vez que entra un nuevo vehículo sin montar en la cadena, aparece un nuevo vehículo montado en la etapa de salida.

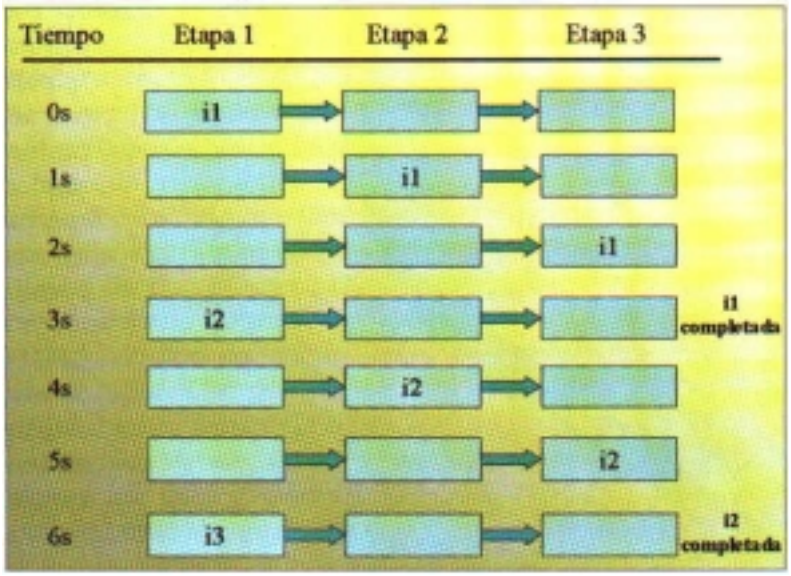


Figura 2. Funcionamiento de un procesador sin pipelining

Suponga que hay N etapas en la cadena, y cada etapa toma 1 segundo en completarse. Si no se emplea la técnica de pipelining, se tarda N segundos en producir cada vehículo. En cambio, si se trabaja con pipelining, se tardará N-1

segundos en llenar las N etapas, pero a partir de ahí cada segundo se obtendrá un coche completo a la salida.

El caso de una CPU es totalmente análogo, teniendo en cuenta que las etapas de la cadena son las etapas de la ejecución de una instrucción. Las Figuras 2 y 3 ilustran el funcionamiento de un procesador con y sin pipelining, empleando 3 etapas de 1 segundo de duración (y, por tanto, trabajando con una frecuencia de reloj de 1 Hz).

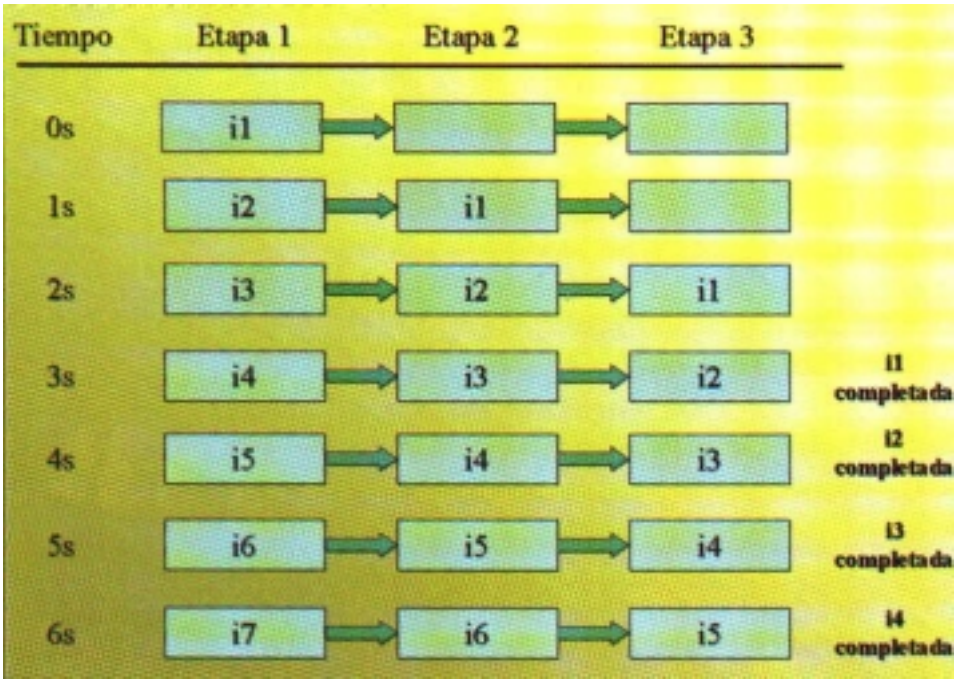


Figura 3. Funcionamiento de un procesador con pipelining

Cuando no se emplea pipelining (Figura 2), se utilizan 3 ciclos de reloj para ejecutar cada instrucción. Si se emplea pipelining (Figura 3), se toman 2 ciclos de reloj para llenar la pipeline, pero a partir de ahí, se ejecuta una instrucción por ciclo de reloj, lo que convierte al procesador en 3 veces más rápido.

Pero, ¿si se implementa una pipeline de N etapas y tras el llenado inicial- realmente obtendremos un procesador N veces más rápido? Como se apreciará en este artículo, esto no es cierto, ya que hay una serie de problemas que rodean al funcionamiento de toda pipeline.

La velocidad de la CPU y la frecuencia de reloj

Conviene hacer un inciso para aclarar un concepto muy importante: ¿cómo comparar la velocidad de dos CPU? La velocidad de ejecución de instrucciones, que se mide en MIPS (millones de instrucciones ejecutadas por segundo), es la cantidad con mayor significado.

Nótese que la velocidad de reloj no se ha nombrado para nada. La frecuencia de reloj relaciona las MIPS y otra cantidad auxiliar: las IPC (número de instrucciones por ciclo de reloj). La operación a efectuar es la siguiente: frecuencia de reloj (MHz)=MIPS/IPC. Por ejemplo, un procesador capaz de ofrecer 100 MIPS y 10 IPC, precisará de una frecuencia de reloj de 100/ 10=10 MHz. En resu-

men, debe recordar que la velocidad real del procesador la dan las MIPS, y no la frecuencia de reloj.

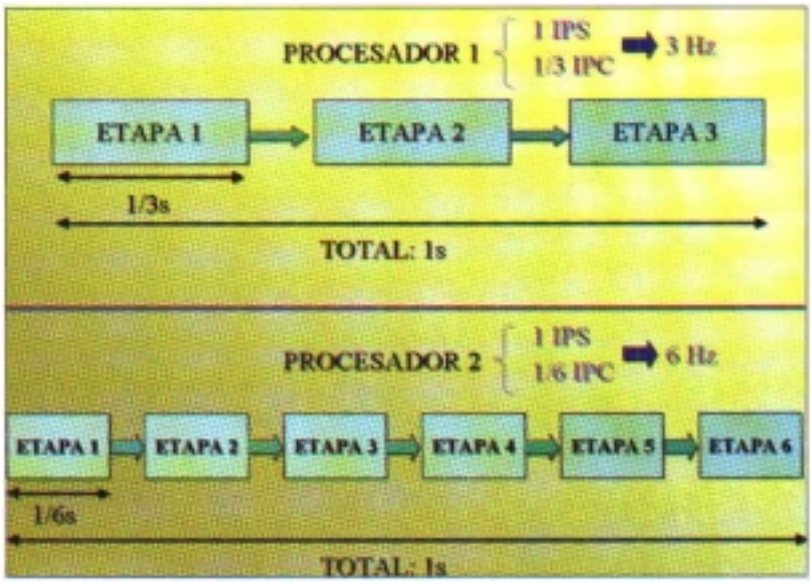


Figura 4. Dos procesadores con idéntica velocidad de procesamiento, pero diferente frecuencia de reloj. La frecuencia de reloj no siempre es un buen parámetro para medir la velocidad de una CPU

Para poner un ejemplo práctico (Figura 4), considere un procesador sin pipelining de tres etapas (procesador 1), donde cada etapa emplea 1/3 segundos en su ejecución. Dicho procesador emplearía 1 segundo (3 etapas x 1/3 segundos cada etapa) en ejecutar cada instrucción, lo que se traduce en 1 instrucción por segundo (1 IPS). Como cada instrucción toma 3 ciclos de reloj, se tiene 1/3 IPC. Por tanto, la frecuencia de reloj necesaria es f reloj=1 IPS/(1/3 IPC)=3 Hz.

Ahora imagine que otro fabricante desglosa cada etapa de la pipeline en 2 nuevas etapas, obteniendo un total de 6 etapas, de 1 / 6 segundos de duración cada una (procesador 2). En este caso, se sigue teniendo 1 IPS pero, en cambio, se obtienen 1/6 IPC. Esto resulta en una frecuencia de reloj de 1 IPS/ (1 / 6 IPC)=6 Hz.

La conclusión es la siguiente: los dos procesadores presentan la misma velocidad de procesamiento (1 IPS), pero la frecuencia de reloj en el segundo caso vale el doble

Por ello, siempre debe recordar la siguiente regla de oro: nunca evalúe ni compare procesadores por su frecuencia de reloj, la cantidad que debe tener en cuenta son las MIPS. Un procesador que trabaja a 200 MHz no tiene por qué ser más potente que otro que trabaja a 100 MHz. En el caso de emplear pipelining, las velocidades se incrementan, pero el concepto presentado es el mismo. La velocidad sólo es útil cuando es la única característica que varía entre dos versiones de un mismo procesador.

Las etapas de la pipeline

Como ya se ha introducido, el pipelining comienza con el desglose en etapas del trabajo de la CPU. La división más básica se reduce a cuatro fases:

Búsqueda de instrucción (Instruction Fetch) Se accede a la memoria para obtener la instrucción a ejecutar.

Decodificación. Se interpreta la instrucción, es decir, se extrae cuál es la operación a realizar y cómo obtener los operandos sobre los que trabajar.

Ejecución. Se buscan los operandos y se realiza la operación indicada por la instrucción.

Almacenamiento de resultados o Write-Back. Se almacena el resultado de la operación en un registro (que posteriormente se puede volcar a la memoria principal). Las CPU actuales implementan muchas más etapas. Por ejemplo, los procesadores Pentium III y los actuales Athlon utilizan del orden de 10 fases. El procesador Pentium 4, en cambio, emplea una pipeline de 20 etapas. Por simplicidad, el resto del capítulo tomará la división en cuatro etapas presentada, ya que representa adecuadamente el funcionamiento de la CPU.

Los siguientes apartados presentan una serie de problemas que rompen con el concepto ideal de pipeline y que, por tanto, perjudican al resultado. Como se podrá comprobar, una pipeline de N etapas incrementa notablemente el rendimiento, pero siempre en un factor menor que N.

El problema de las burbujas

La selección de las etapas a implementar es una decisión delicada para los diseñadores de una CPU. El primer detalle a tener en cuenta es el equilibrio entre etapas. Estas deberían ser capaces de ejecutarse en el mismo número de ciclos de reloj. En caso contrario, se producirían "atascos": mientras la etapa más lenta termina su trabajo, la cadena queda bloqueada. Por ello se puede decir que la etapa más lenta condiciona el rendimiento de la pipeline, y por tanto hay que igualar tiempos a toda costa buscando una división óptima.

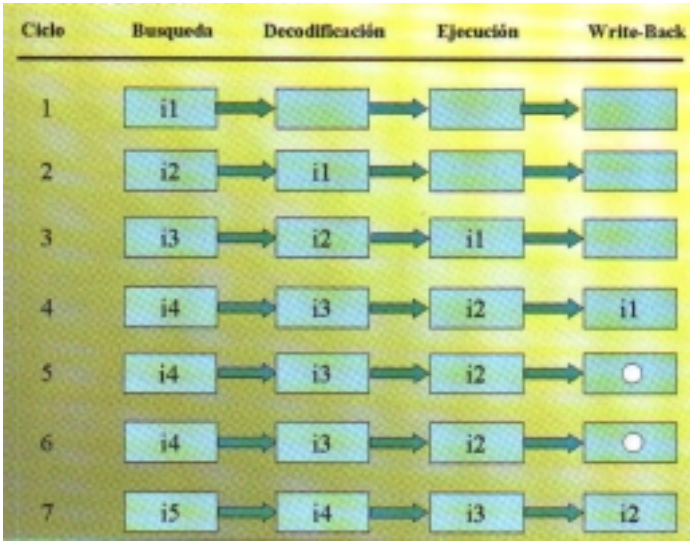


Figura 5. Formación de burbujas en la pipeline

Este problema no es fácil de solucionar. A pesar de realizar un desglose en etapas suficientemente equilibrado, hay etapas que -en determinadas circunstancias- toman más de un ciclo de reloj. Tal es el caso de instrucciones complejas, como la multiplicación y la división, que prolongan la etapa de ejecución. Lo mismo sucede con la lectura desde la memoria en la fase de búsqueda de instrucción y ejecución. La escritura en memoria (en la fase de almacenamiento de resultados) también puede tomar más de un ciclo de reloj.

Cuando una etapa toma más tiempo del esperado (entorpeciendo el avance normal de la pipeline) se dice que se ha formado una burbuja en la etapa siguiente. Este concepto se aprecia claramente en la Figura 5, donde la instrucción **i2** toma tres ciclos de reloj en la etapa de ejecución, paralizando el avance de la pipeline durante ese tiempo. Se ha producido una burbuja en la última fase.

El problema de las burbujas complica más el diseño, ya que se hace necesario un mecanismo de sincronización, que asegure que los datos pasan de una etapa a la siguiente cuando esta última quede libre.

El problema de los saltos

Uno de los problemas con más impacto en el rendimiento de la pipeline surge ante la ejecución de instrucciones de salto. Tome como referencia la pipeline de la Figura 6. Imagine que **i2** es una instrucción que produce un salto a cierta instrucción **ia**. Cuando **i2** llega a la fase de ejecución, las etapas anteriores se han ido llenando con las instrucciones que le suceden, en orden lógico. Pero tras ejecutar **i2**, la siguiente instrucción que debe pasar por la fase de ejecución... ¡no es **i3**, sino **ia**! Esto obliga a vaciar las etapas anteriores, creando burbujas hasta que la instrucción **ia** se localiza en la memoria. La pipeline recupera su funcionamiento normal cuando **ia** llega a la fase de ejecución, lo que supone un retardo de 3 ciclos.

La gravedad de este problema radica en una estadística bastante acertada: un 10% de las operaciones que ejecuta un procesador son instrucciones de salto. Por tanto, el rendimiento de la pipeline va a ser negativamente afectado con demasiada frecuencia.

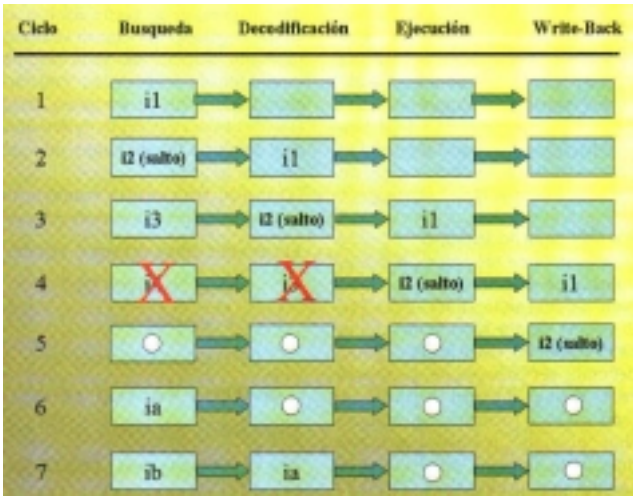


Figura 6. Problema de las instrucciones de salto

Una solución se halla en los compiladores (el software que genera las instrucciones que entiende la máquina a partir de instrucciones de alto nivel). Al generar el código a ejecutar, el compilador puede reordenar las instrucciones, colocando la instrucción que precede al salto después de este. El orden resultante es: **i2** (salto), **i1**, **i3**, **i4**. De esta forma, cuando **i2** llega a la fase de ejecución, actualiza el contador de programa, apuntando a **ia** en lugar de **i3**. Por tanto, **ia** pasa a la primera etapa. De esta manera no hay necesidad de vaciar la pipeline, ya que las instrucciones se ejecutan en el orden adecuado, y no entra ninguna instrucción no deseada. Todo ello queda ilustrado en la Figura 7.

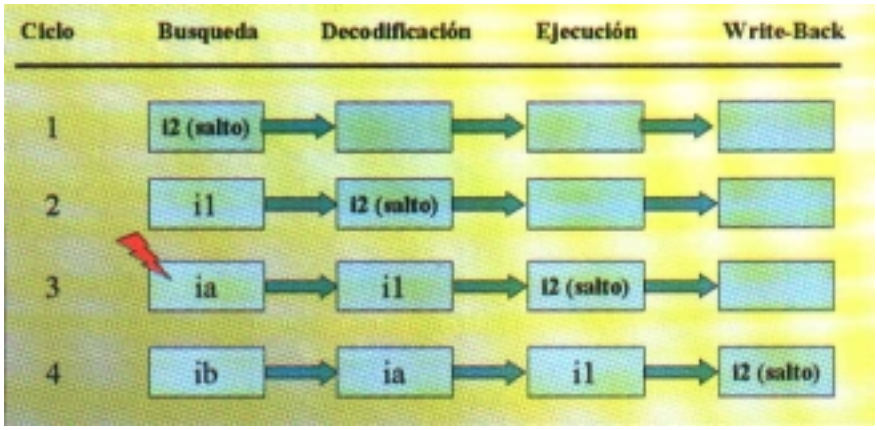


Figura 7. Solución al problema de las instrucciones de salto mediante reordenamiento (via software)

Los compiladores son más potentes cada día en este aspecto. Reordenar el código de forma inteligente reduce considerablemente la necesidad de vaciar la pipeline ante situaciones como los saltos, lo que lleva a un notable aumento del rendimiento de la CPU.

Otra solución consiste en reestructurar el código a la hora de compilar, intentando evitar los saltos a toda costa. Por ejemplo, se puede deshacer un bucle de N etapas en un código repetitivo N veces más voluminoso, pero que se ejecuta de forma secuencial.

Saltos condicionales y dependencia de instrucciones

Todavía quedan dos problemas importantes por resolver. El primero radica en la existencia de saltos condicionales en los programas, algo que sucede con frecuencia. Por ejemplo, considere una instrucción *JNZ dirección_destino*, que salta a la dirección indicada si la última instrucción ha devuelto un valor no nulo como resultado (continuando normalmente el programa en caso contrario). Cuando dicha instrucción llegue a la fase de ejecución, la CPU aún no conoce cuál será el destino del salto. En efecto, la instrucción anterior al salto se encuentra en la fase de almacenamiento de resultados, y sólo después de terminar dicha etapa se conocerá si el resultado ha sido nulo. Si el resultado es nulo, la pipeline puede continuar su evolución normal. En caso contrario; se produce un salto, y por tanto aparecerán burbujas al tener que vaciar la pipeline. Nótese que, en este caso, un compilador no conoce de antemano cuándo se va a producir un salto, y no es posible realizar reordenaciones como las comentadas anteriormente.

El segundo problema se basa en la dependencia entre instrucciones.

Sea el siguiente fragmento de código (trabajando en un lenguaje de alto nivel):

(1) c=a+b

(2) Si c>4, entonces d=2*c

(3) Si c<4, entonces d=-10*c

La instrucción 1 se ejecuta sin problemas, pero la siguiente instrucción depende del resultado de sumar las variables a y b. La CPU tampoco sabe de antemano cuál será el resultado de la suma, y por tanto habrá ocasiones en que se precise de un vaciado de la pipeline (concretamente, cuando c es menor que 4).

Ambos problemas se reducen a una única pregunta: ¿cuál será la instrucción siguiente al salto? Una solución dicha pregunta -esta vez basada en hardware- es la implementación de una unidad de predicción de saltos. Se trata de un subsistema electrónico integrado en la CPU, que realiza una predicción de los resultados que motivan este tipo de saltos (por supuesto, la predicción se basa en el estudio del comportamiento de los programas por parte del procesador). En el código anterior, el procesador tratará de predecir el valor de la variable c. Si dicha variable ha estado tomando el valor 10 en las últimas 20 ocasiones, la predicción evidente será "c=10". La CPU confía en su predicción y continúa de forma normal. En el caso de la predicción contraria (c<4) la CPU evitaría el conflicto en la pipeline con antelación, entendiendo que **i3** es la instrucción que sucede a **i1**.

Como puede imaginar, la unidad de predicción de saltos no siempre acertará, y en esos casos la pipeline se llenará de instrucciones que no siguen el orden adecuado. Esto último implica un retardo para vaciar la pipeline, y afecta negativamente al rendimiento del procesador. Afortunadamente, la mayoría de los programas suele contener bloques de código repetitivos, lo que permite obtener un 90% de las predicciones correctas en la mayoría de los casos.

¿Cuántas etapas?

Al iniciar este artículo, la idea que parecía reinar era dividir la pipeline en el mayor número de etapas posible. Tan sólo seria necesario tener en mente algunos principios, como igualar el tiempo empleado en cada etapa para evitar la formación de burbujas. Esto conseguiría incrementar el rendimiento de la CPU (y la frecuencia de reloj) si la pipeline fuera ideal, es decir, si no existieran efectos tan amenazantes como los fallos de predicción de saltos. Desgraciadamente, dichos problemas son inevitables, y condicionan el diseño en gran medida. Por ejemplo, sea el caso del procesador Pentium 4 que se ha diseñado con una pipeline de 20 etapas. Si ocurre una falsa predicción, serán necesarios 19 ciclos de reloj para el vaciado de la pipeline. En un procesador con menor número de etapas (por ejemplo, Pentium III) el impacto en el rendimiento es menor, pero la frecuencia de reloj a emplear es también menor (y desgraciadamente es el factor con mayor impacto comercial). Algunos estudios anteriores al lanzamiento del Pentium 4 estimaban que éste sería un 20% más lento que el Pentium III debido a la excesiva longitud de su pipeline.

Al incrementar el número de etapas surge otro problema: el espacio físico necesario. A mayor número de etapas, mayores dimensiones requiere la CPU. Sirvan como prueba las dimensiones del chip de los procesadores Pentium 4, que son aproximadamente el doble que las dimensiones del Pentium III.

CAPÍTULO 4

LA UNIDAD CENTRAL DE PROCESO (III)

Tras haber estudiado con detalle la estructura de una CPU sencilla y una técnica de gran importancia (el pipelining), disponemos de los ingredientes fundamentales para comprender el funcionamiento de cualquier CPU moderna. No obstante, la arquitectura de los procesadores actuales es mucho más compleja que la que hemos presentado. Por ello, es conveniente mostrar algunos ejemplos reales para tener una idea clara de cómo se llevan a la práctica los conceptos abordados hasta la fecha. Este artículo se centra en dos procesadores de la firma Intel: Pentium III y Pentium 4. Su conocimiento no sólo refuerza al máximo el aprendizaje de la CPU, sino que proporciona información técnica valiosa acerca de dos procesadores muy actuales y presentes en la mayoría de PC de todo el mundo. En todo momento apreciará cómo, a pesar de tratarse de sistemas complejos, los ingredientes presentados en las anteriores entregas se hallan siempre presentes en el diseño y funcionamiento de estas CPU.

El procesador Pentium III

Antes de abordar el procesador Intel Pentium 4 es conveniente tener una idea de la estructura del procesador Pentium III, perteneciente a la generación inmediatamente anterior. Como se apreciará más adelante, el procesador Pentium 4 contiene un esqueleto similar al del Pentium III, aportando una serie de novedades que -en principio- contribuyen a una mejora del rendimiento.

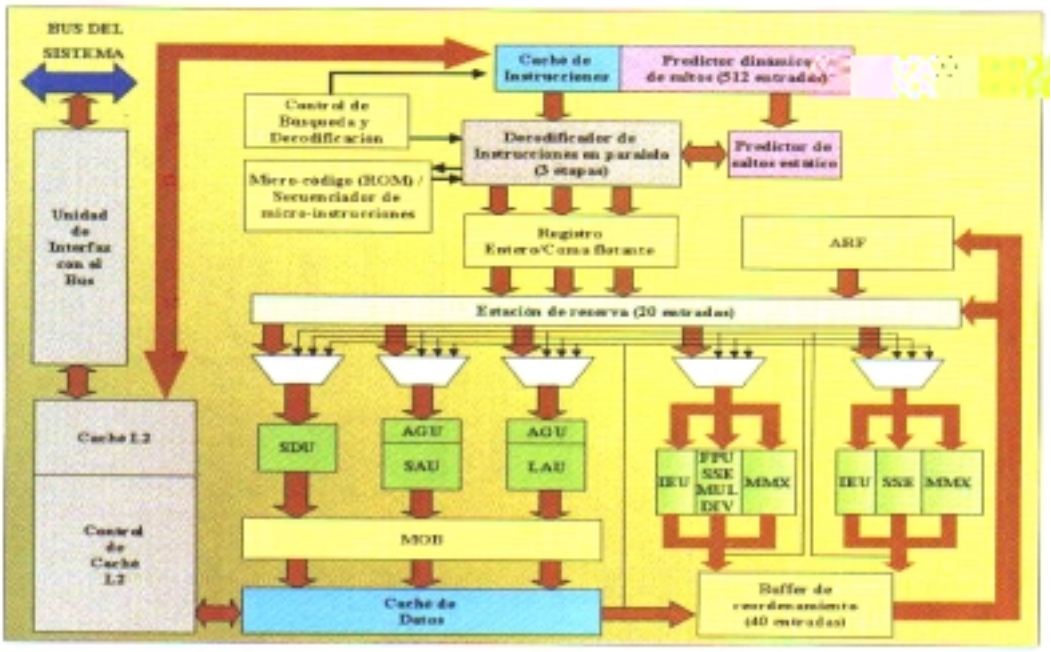


Figura 1. Arquitectura del procesador Pentium III.

La Figura 1 resume la estructura interna del procesador Pentium III en forma de diagrama de bloques. Se pueden apreciar algunas etapas comentadas en las anteriores entregas de esta serie. En concreto -y hablando a alto nivel- se distinguen grupos de bloques que se corresponden con las 4 etapas más básicas de una pipeline (búsqueda de instrucción, decodificación, ejecución y almacenamiento de resultados).

En primer lugar, se aprecia una etapa de *búsqueda de instrucciones* y una unidad de *decodificación* (que en realidad son tres, trabajando en paralelo).

Bajo la etapa denominada "estación de reserva" se alojan varias unidades de *ejecución* especializadas. Algunas de ellas se dedican a las operaciones con enteros, y otras a las operaciones en coma flotante, operaciones con direcciones de memoria, etc. También se distinguen unidades dedicadas a operaciones complejas, como la multiplicación y la división, y a otros grupos de instrucciones (como las extensiones multimedia MMX). Llama la atención una unidad de procesamiento dedicada al grupo de instrucciones SSE *(Streaming SIMD Extension)*. Estas instrucciones son de tipo SIMD *(Single Instruction Multiple Data)*, lo que significa que permiten ejecutar una misma instrucción sobre un conjunto de datos, lo cual proporciona muchas ventajas en aplicaciones científicas, multimedia, de ingeniería, etc. Por supuesto, en esta etapa destacan las ALU (introducidas en anteriores capítulos). En el caso de las operaciones con direcciones de memoria, el elemento destacado recibe el nombre de AGU *(Address Generation Unit*, unidad de generación de direcciones).

Finalmente, en la parte inferior del esquema aparecen los bloques dedicados al almacenamiento de resultados. Nótese la presencia de conexiones que conducen los resultados de nuevo hacia las unidades de procesamiento, lo que recuerda totalmente la estructura básica de una CPU, tal y como la presentamos en el segundo capítulo.

Por supuesto, no pueden faltar elementos ya conocidos, como las unidades de predicción de saltos, las memorias caché, los registros y la unidad de interfaz con el bus.

En cuanto a la memoria caché, se debe notar que está dividida en dos bloques. En primer lugar, aparece la caché denominada "de nivel 1" o "L1". Ésta presenta un reducido tamaño y el acceso es rápido, siendo la primera que se consulta. Si la información no se encuentra allí, se recurre a la caché "de nivel 2" o "L2", que presenta un tamaño mayor. El acceso es más lento, pero sigue siendo mucho más rápido que acceder a la memoria principal. El procesador Pentium III presenta 2 cachés de tipo L1 (una dedicada a datos y otra a instrucciones) de 16 kB, y una memoria caché de tipo L2 de 256 kB, donde se comparten instrucciones y datos. Se puede apreciar claramente la conexión existente entre las memorias caché L1 y caché L2, lo que permite el funcionamiento en dos niveles.

Otro elemento interesante es el secuenciador de micro-instrucciones, acompañado de una memoria de sólo lectura (ROM). Como ya comentamos, conviene desglosar las instrucciones más complejas en una secuencia de instrucciones que trabajan a nivel interno (denominadas micro-operaciones). Tan sumamente bajo es su nivel de operación, que los bits que conforman estas instrucciones operan directamente sobre elementos internos de la CPU. Cada secuencia de

micro-instrucciones compone lo que se denomina un micro-programa. En la memoria ROM se almacenan los micro-programas relacionados con algunas instrucciones de alto nivel de complejidad.

No es el objetivo de esta descripción convertir al lector en un experto en las interioridades de cada bloque interno a la CPU. En cambio, hemos ofrecido una visión general que permite observar cómo incorpora un procesador moderno y potente los elementos y. técnicas presentados en las anteriores entregas. Los próximos apartados muestran una visión similar sobre el último procesador creado por Intel: el Pentium 4.

Introducción al Pentium 4

El nuevo procesador Pentium 4 aparece cubierto por un nuevo concepto: la micro-arquitectura NetBurst de Intel. Dicha arquitectura, según afirma el fabricante, proporciona una considerable ganancia en rendimiento en multitud de áreas de aplicación.

Dicho procesador fue introducido en noviembre de 2000 con una velocidad de reloj de 1,5 GHz. Según asegura Intel, la arquitectura NetBurst aporta un gran incremento de rendimiento en aplicaciones como la transmisión de audio y vídeo a través de Internet, procesamiento de imágenes, reconocimiento de voz, videoconferencia, etc.

El procesador se compone de 42 millones de transistores, con 6 niveles de interconexiones de aluminio. El tamaño del chip es de 217 mm^2 y consume 55 vatios de potencia a 1,5 GHz. El bus de sistema permite trabajar a la llamativa velocidad de 3,2 GB/s (gracias al empleo de un tipo especial de memoria - particularmente cara denominada RDRAM). Esto proporciona un claro beneficio en el rendimiento del procesador.

Otra característica interesante es la adición de 144 nuevas instrucciones de 128 bits, englobadas bajo el acrónimo SSE2 (Streaming SIMD Extensión 2), que constituyen una evolución de las instrucciones SSE presentes en el Pentium III. Según afirma Intel, estas instrucciones mejoran el rendimiento en aplicaciones multimedia, científicas y de ingeniería.

Es un buen momento para indicar un detalle importante: es sencillo encontrar multitud de documentos técnicos en Internet que demuestran que el rendimiento del procesador Pentium 4 no es tan espectacular como afirma el fabricante, en multitud de áreas de aplicación. Este trabajo no persigue realizar una valoración del rendimiento del Pentium 4 ni tampoco una comparativa con otros procesadores que parecen ofrecer mejores resultados. Tan sólo pretendemos mostrar un recorrido superficial a través de los componentes fundamentales de la CPU. Los conocimientos de fondo adquiridos hasta ahora se pueden complementar con infinidad de artículos tipo "banco de pruebas" o "benchmark". De esa forma, es posible sacar conclusiones sobre la información que proporcionan los fabricantes, analizando la exactitud de sus afirmaciones.

Los siguientes apartados enfocan más de cerca lo que se esconde en el interior del procesador Pentium 4, sin entrar en los detalles más internos de cada bloque funcional (lo que conduciría, inevitablemente, al mundo de la electrónica

digital y otras disciplinas relacionadas). Se recomienda consultar la Figura 1 para apreciar las similitudes con la estructura del procesador Pentium III.

La arquitectura NetBurst

La Figura 2 resume la arquitectura interna del procesador Pentium 4 (NetBurst) mediante un diagrama de bloques. Se aprecian cuatro bloques principales:

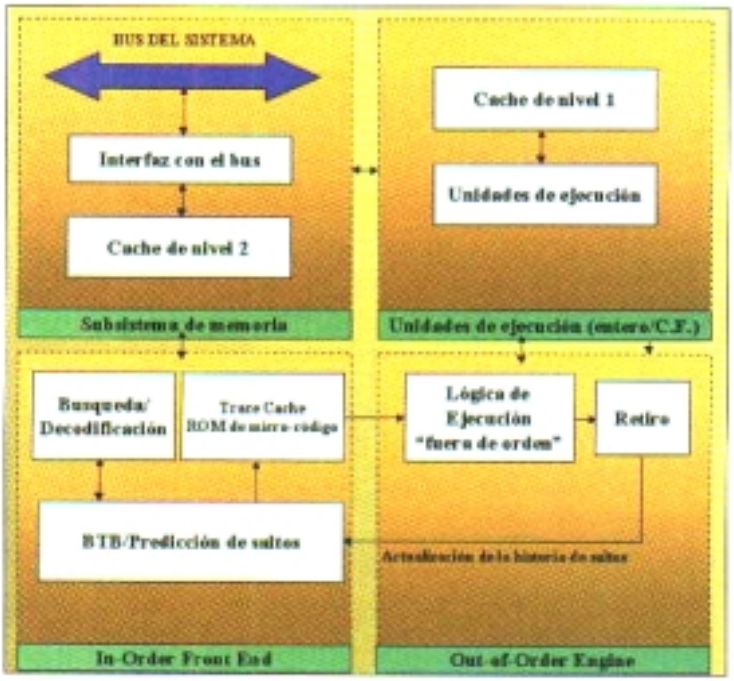


Figura 2. Diagrama de bloques general de la arquitectura NetBurst

- *In-order front end*. Esta etapa se encarga de buscar las instrucciones que los programas van a emplear próximamente, y las prepara para ser utilizadas en la pipeline. El objetivo es proporcionar un flujo de instrucciones decodificadas a gran velocidad, de forma que sólo falte completar su ejecución. Se incluye un potente bloque de predicción de saltos, que emplea la información recogida durante la ejecución de las anteriores instrucciones para predecir cuáles serán las siguientes instrucciones a ejecutar. De esta forma, dichas instrucciones se pueden buscar en la caché L2 con antelación, confiando en las predicciones (el porcentaje de errores es realmente reducido). Las instrucciones se decodifican y se transforman en micro-operaciones, que son las instrucciones que el procesador realmente comprende (ya que trabajan directamente sobre los componentes internos). El bloque denominado Trace Cache es una versión avanzada de la caché L1 presente en el procesador Pentium III. La innovación radica en que la caché se conecta directamente al motor de ejecución del sistema (que sólo entiende micro-operaciones, y no instrucciones sin decodificar). En otras palabras, la Trace Cache permite almacenar las instrucciones previamente buscadas y decodificadas (léase, convertidas a micro-operaciones). La forma de trabajar es análoga a la que se tenía en el caso del Pentium III, con la diferencia de que se trabaja con instrucciones de nivel más bajo.

- *Out-of-Order Engine* (motor de ejecución fuera de orden). Este bloque se encarga de planificar las micro-operaciones (en adelante, simplemente denominadas instrucciones) para su ejecución. La lógica de ejecución fuera de orden incorpora varias etapas de almacenamiento (buffers) que reordenan el flujo de instrucciones, haciéndolo óptimo para un buen funcionamiento durante su avance por la pipeline. El reordenamiento permite que las instrucciones se ejecuten tan pronto como se disponga de los operandos necesarios (evitando el retardo que produce dicha espera, lo que ya se bautizó como burbujas en la pipeline). Las instrucciones que vienen a continuación de una instrucción que produce retardos -y que no dependen de esta última- se ejecutan con anterioridad. De esta forma, cuando dichas instrucciones acaban su ejecución, la instrucción en espera ya dispone de los operandos que buscaba, y todo continúa de forma fluida. En resumen, la ejecución fuera de orden persigue que los recursos implicados en la ejecución (por ejemplo, las ALU y la caché) permanezcan ocupados en la mayor medida posible en todo instante, buscando el mayor rendimiento posible.

La etapa restante (lógica de retiro) se encarga de volver a colocar las instrucciones en su orden original. El procesador Pentium 4 es capaz de retirar 3 instrucciones por ciclo de reloj. Nótese que este bloque conoce las últimas operaciones de salto realizadas. Por ello, también se encarga de comunicar dicha información al predictor de saltos, de forma que se pueda entrenar con la última información real conocida.

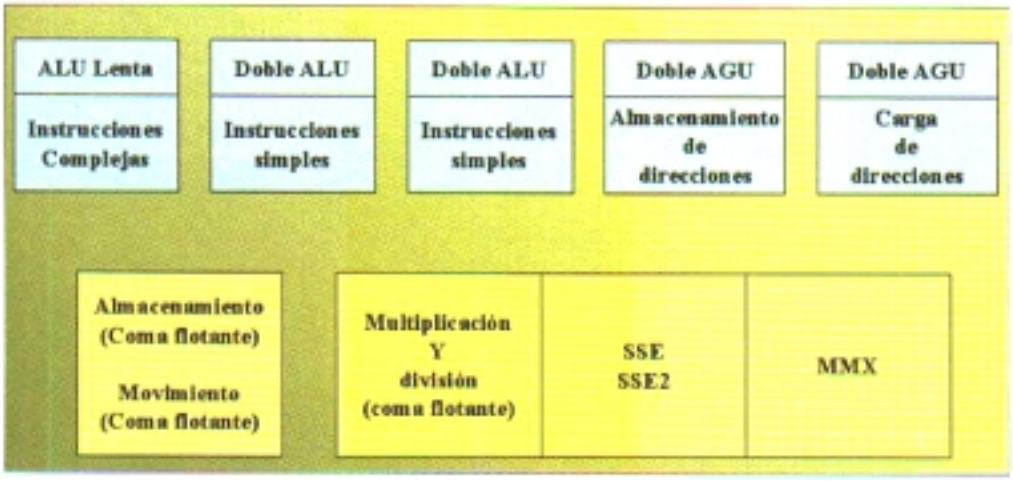


Figura 3. Unidades de ejecución del procesador Pentium 4

- *Unidades de ejecución (con enteros y en coma flotante).* Este bloque contiene las etapas donde se hace efectiva la ejecución de instrucciones. El bloque denominado unidades de ejecución es el que contiene los elementos de procesamiento, especializados en multitud de operaciones con datos enteros y en coma flotante. La Figura 3 muestra la estructura interna de dicha etapa. Se puede detectar la presencia de varios bloques ya comentados en este artículo y en los anteriores (sobre todo en lo referente a las ALU).

Por supuesto, tras la ejecución de las instrucciones, se requiere pasar a la fase de almacenamiento de resultados, de ahí la incorporación de la memoria caché L1.

- *Subsistema de memoria*. En este bloque se incluye, por supuesto, la memoria caché de nivel 2. Como ya hemos introducido, esta memoria almacena las instrucciones y datos que no se encuentran en el bloque Trace Cache y cuyo acceso será necesario con menor frecuencia. Como es de esperar, en ocasiones se intentará acceder a un dato o instrucción que no se encuentra en la memoria caché de nivel 2. Por ello, ésta se encuentra conectada al bus del sistema (a través de la unidad de interfaz con el bus), de forma que pueda acceder a la memoria principal (y unidades de E/S) para buscar dicha información. Es importante remarcar la división de la memoria en 3 niveles (memoria principal-caché L2-caché L1). Cada nivel tiene una capacidad inversamente proporcional a la probabilidad de encontrar la información deseada y a la velocidad de acceso. La memoria principal es la que se accederá el menor número de veces (ya que la información se encontrará en las cachés Ll y L2 en la mayoría de las ocasiones), y por ello presenta una gran capacidad de almacenamiento, y un acceso lento. En cambio, en la caché L1 se encontrará la información deseada la mayoría de las veces, y por ello presenta una capacidad y velocidad de acceso reducidas.

Un poco más de detalle

La Figura 4 muestra un diagrama de bloques de la arquitectura NetBurst equivalente al mostrado en la Figura 2, pero aportando algo más de detalle. En el diagrama se incluye un desglose de las etapas comentadas en el apartado anterior. En la parte izquierda se sigue apreciando el subsistema de memoria tal y como se ha introducido. Los bloques de la parte superior no son más que un desglose de los elementos que componen la etapa In-Order Front End. En la parte central se ha desglosado el motor de ejecución fuera de orden. Finalmente, en la parte inferior se aprecia el motor de ejecución y la caché de nivel 1, que forman el bloque de unidades de ejecución introducido anteriormente.

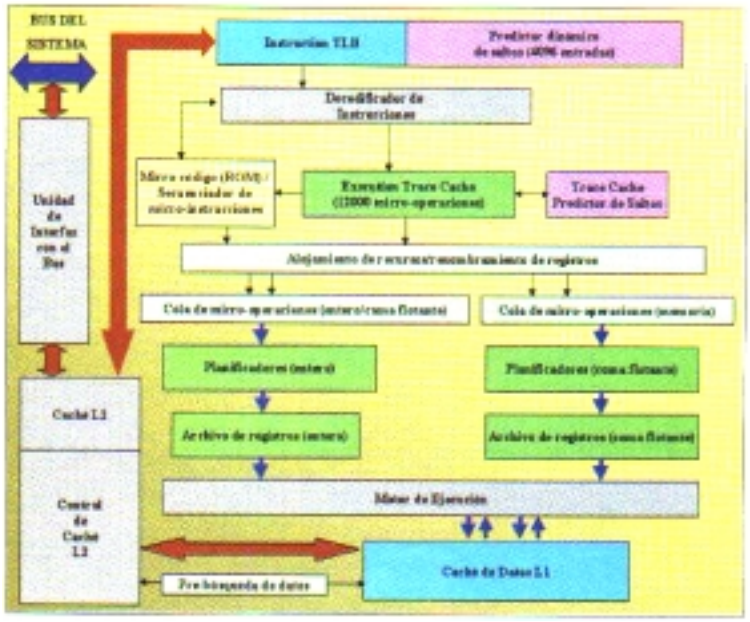


Figura 4. Arquitectura del procesador Pentium 4 (arquitectura NetBurst, con mayor detalle)

Es interesante remarcar la característica de prebúsqueda de datos *(prefetch)* en la caché L2. Esta característica permite buscar datos en la caché antes de que sean requeridos para la ejecución de instrucciones, lo que aporta una mejora de rendimiento importante. En el procesador Pentium III, esta característica se implementaba mediante instrucciones SSE. En el nuevo Pentium 4 se realiza mediante hardware, lo que aumenta aún más la velocidad de dicha etapa.

Otro bloque que no aparecía en la Figura 2 es el de alojamiento de recursos. Éste se encarga de asignar los recursos (registros, etc.) que necesitan las micro-operaciones de forma efectiva (tal y como se indicaba, se trata de intentar mantener los recursos ocupados al máximo durante todo el tiempo).

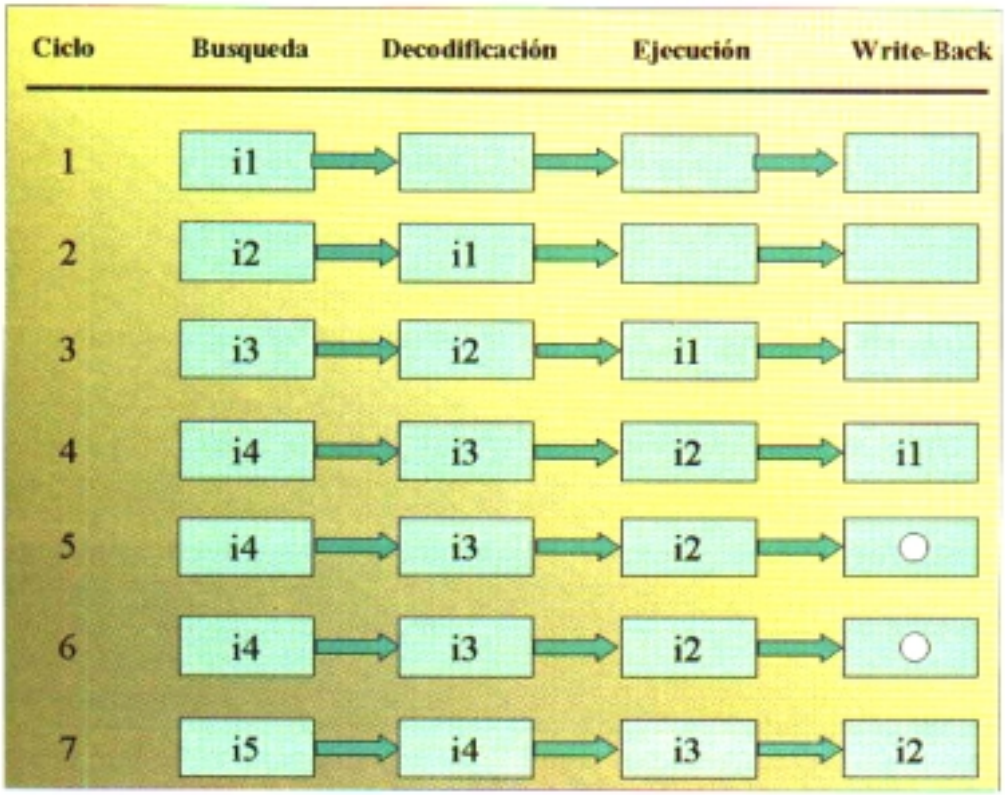


Figura 5. La pipeline del procesador Pentium 4

Otra etapa que complementa a la anterior es el renombramiento de registros. En este caso, se trata de asegurar la compatibilidad con el código procedente de anteriores generaciones del procesador, que contenían un archivo de registros más pequeño (el procesador Pentium 4 contiene 128 registros en su archivo).

Un elemento importantísimo son los planificadores de micro-operaciones. Estos constituyen el corazón del motor de ejecución fuera de orden, y son los responsables del reordenamiento de instrucciones antes comentado. Las micro-operaciones a reordenar proceden de dos colas, una especializada en las operaciones sobre datos enteros o en coma flotante, y otra dedicada a las operaciones sobre la memoria. Finalmente, no podían faltar los registros, que se diferencian según el tipo de datos con el que trabajar (datos enteros o en coma flotante).

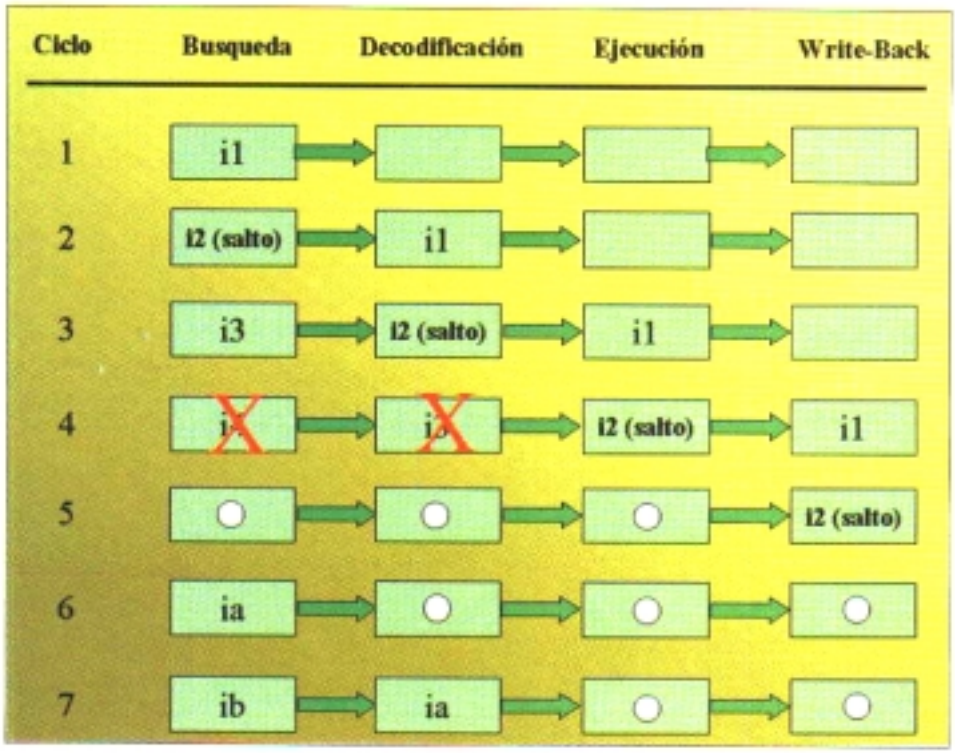


Figura 6. La pipeline del procesador Pentium III

Para concluir el presente recorrido, es conveniente dar un vistazo de alto nivel al funcionamiento del procesador, lo que apunta hacia la pipeline. La Figura 5 muestra las 20 etapas que componen la pipeline del procesador Pentium 4. En la Figura 6 se muestra la pipeline del procesador Pentium III, que contiene la mitad de etapas. Como se estudió en la anterior entrega, el hecho de duplicar el tamaño de la pipeline permite incrementar la frecuencia de reloj, pero no tiene por qué ser un indicador de que el rendimiento es mayor. Muchos expertos opinan que el diseño de dicha pipeline procede del departamento comercial de Intel, ya que la frecuencia de reloj es el factor que "vende". De hecho, se pueden encontrar estudios que afirman que el procesador Pentium 4 presenta peor rendimiento que el Pentium III en varios tipos de aplicaciones. Como ya se comentó, la longitud de la pipeline también impacta en otros factores, como el tamaño del chip (que en el caso del Pentium 4 se hace notar).

CAPÍTULO 5

LA MEMORIA PRINCIPAL

En los anteriores capítulos, queda claro que la CPU es el corazón del PC. En efecto, es el motor que se encarga de ejecutar las tareas que indican las instrucciones. Pero las instrucciones, y los operandos sobre los que éstas trabajan, se deben tomar de algún almacén de información de rápido acceso. Lo mismo ocurre con los resultados que producen las instrucciones, que pueden servir en instantes posteriores y por tanto deben ser almacenados en algún lugar rápidamente accesible. Este artículo se centra en el elemento que resuelve dicho problema, un subsistema imprescindible para que la CPU pueda trabajar: la memoria RAM. Se introducirán sus características, la razón de su existencia, su estructura interna, su apariencia física, las tecnologías existentes y las técnicas para la verificación de la integridad de los datos.

Características de la memoria principal (RAM)

Un sistema de memoria se puede clasificar en función de muy diversas características. Entre ellas se pueden destacar las siguientes: localización de la memoria, capacidad, método de acceso y velocidad de acceso. En el caso de la memoria RAM *(Random Access Memory*, también denominada memoria principal o primaria), se puede realizar la siguiente clasificación:

Localización. Interna (se encuentra en la placa base).

Capacidad. Hoy en día, no es raro encontrar ordenadores PC equipados con 64 ó 128 MB de memoria RAM.

Método de acceso. La RAM es una memoria de acceso aleatorio. Esto significa que una palabra o byte se puede encontrar de forma directa, sin tener en cuenta los bytes almacenados antes o después de dicha palabra (al contrario que las memorias en cinta, que requieren de un acceso secuencial). Además, la RAM permite el acceso para lectura y escritura de información.

Velocidad de acceso. Actualmente se pueden encontrar sistemas de memoria RAM capaces de realizar transferencias a frecuencias del orden de los Gbps (gigabits por segundo).

También es importante anotar que la RAM es una memoria volátil, es decir, requiere de alimentación eléctrica para mantener la información. En otras palabras, la RAM pierde toda la información al desconectar el ordenador.

¿Para qué sirve la memoria RAM?

Una vez introducidas sus características, la pregunta inmediata que surge es: ¿qué sentido tiene utilizar la RAM (sobre todo, teniendo en cuenta que es volátil)? Evidentemente, los programas se deben almacenar en unidades capaces

de mantener la información sin necesidad de alimentación eléctrica, ya que el objetivo es volverlos a utilizar en futuras sesiones. De la misma forma, también es deseable almacenar datos con los que trabajarán dichos programas (por ejemplo, información de configuración). Las unidades de almacenamiento típicamente empleadas para ello son los discos duros.

Imagine que la CPU tuviera que acudir al disco duro para tomar cada instrucción de un programa, tomar los datos que pueda necesitar, escribir resultados intermedios o finales, etc. Afortunadamente, los discos duros son unidades con gran capacidad de almacenamiento, lo que es muy apropiado para almacenar grandes cantidades de instrucciones y datos. Pero; lamentablemente, dichas unidades se caracterizan por un lento acceso a la información, por lo cual la CPU ejecutaría los programas con una velocidad mucho menor de la que es capaz de ofrecer. En cambio, a la RAM se accede de forma rápida (y por tanto es apropiada para trabajar "en equipo" con la CPU), aunque no dispone de tanta capacidad de almacenamiento y es volátil.

Por tanto, la solución ideal es la combinación de ambas memorias. Los programas y datos se leen desde el disco duro, y se copian en la memoria principal, de forma que la ejecución sea eficiente. Los resultados intermedios se leen y escriben también en la RAM. Cuando un programa finaliza, los resultados que es necesario almacenar se guardan en el disco duro, de forma que están disponibles de cara a futuras sesiones. Un símil interesante es pensar en un despacho de oficina. Los archivadores de documentos hacen el papel del disco duro, y el escritorio hace el papel de la memoria RAM. A la hora de trabajar, se sacan del archivador los documentos que se van a emplear, y se colocan en el escritorio para tener un rápido acceso a los mismos. Tras el trabajo, los documentos generados como resultado se guardan de nuevo en el archivador, de forma que estén disponibles para futuras jornadas de trabajo. La desconexión del ordenador equivaldría a tirar los documentos del escritorio a la papelera (se perdería la información). Si se trabajara directamente sobre el archivador, sería necesario abandonar el escritorio y caminar para tomar o dejar documentos, lo que daría lugar a un trabajo poco eficiente.

Otra posible cuestión es: ¿qué ocurre si un programa no cabe en la memoria? En ese caso, es necesario aplicar algunas técnicas que se comentarán en futuros capítulos. Lo que sí resulta evidente es que a mayor cantidad de memoria, mayor número de programas se pueden abrir simultáneamente.

Apariencia física

La memoria RAM se presenta en forma de circuitos integrados (chips). El interior de cada chip se puede imaginar como una matriz o tabla, en la cual cada celda es capaz de almacenar un bit. Por tanto, un bit se puede localizar directamente proporcionando una fila y una columna de la tabla. En realidad, la CPU identifica cada celda mediante un número, denominado dirección de memoria. A partir de una dirección, se calcula cuál es la fila y columna correspondiente, con lo que ya se puede acceder a la celda deseada. El acceso se realiza en dos pasos: primero se comunica la fila y después la columna, empleando los mismos terminales de conexión. Obviamente, esta técnica -denominada multiplexado, permite emplear menos terminales de conexión para acceder a la

RAM, lo que optimiza la relación entre el tamaño del chip y la capacidad de almacenamiento.

Realmente, la CPU no suele trabajar con bits independientes, sino más bien con agrupaciones de los mismos, en forma de palabras binarias. Esto hace que la RAM no se presente en un solo chip, sino más bien en agrupaciones de los mismos. Por ejemplo, un grupo de 8 chips, cada uno capaz de almacenar N bits, proporcionará en conjunto N kB. Antiguamente, la RAM se soldaba directamente a la placa base del ordenador. Esto tenía varios inconvenientes: era difícil reemplazar los chips o ampliar su número, y además se ocupaba demasiado espacio en la placa base. Actualmente, estos problemas se han solucionado, gracias a la presentación de la memoria en dos formatos estándar: SIMM *(Single In-line Memory Module)* y DIMM *(Dual In-line Memory Module)*. Estos formatos contienen chips de memoria soldados en placas de circuito impreso. Por supuesto, existe una interfaz de conexionado estándar con la placa base, a través de unas ranuras de conexión instaladas sobre ella. Los módulos se insertan en posición vertical o inclinada, lo que optimiza la ocupación de espacio en la placa base. Además, la existencia de varias ranuras de conexión permite ampliar la memoria o reducirla de forma sencilla. Existen otros formatos de presentación de la memoria, pero en el mundo del PC, los acrónimos SIMM y DIMM destacan por encima del resto.

Aún hay un nivel de jerarquía más alto: los bancos de memoria. En efecto, los módulos se agrupan formando bancos, a los cuales se dirige la CPU (de forma no simultánea). El número de bancos y los conectores incluidos en cada uno de ellos son datos determinados por la CPU y por la forma en que ésta envía y recibe la información (el número de bytes que la CPU puede procesar simultáneamente).

Figura 1. Módulo SIMM de 30 contactos

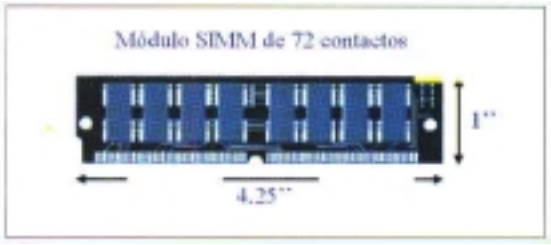


Figura 2. Módulo SIMM de 72 contactos

Módulos SIMM

Las Figuras 1 y 2 muestran el aspecto de los módulos SIMM existentes. En primer lugar, se encuentran los módulos de 30 contactos (el primer tipo de SIMM que apareció, característico de los PC más antiguos). En sus diversas configuraciones, estos módulos son capaces de ofrecer 1, 2, 4 ó 16 MB. Además, son capaces de transferir 8 bits de datos. Una CPU de 32 bits requiere de 4 módulos SIMM de 30 contactos en cada banco de memoria, para poder proporcionar 32 bits.

Después aparecieron los módulos de 72 contactos, como respuesta a los crecientes requisitos de memoria de los ordenadores personales. Estos módulos son capaces de transferir 32 bits de datos, y presentan configuraciones de 4, 8, 16, 32, 64 MB, etc. Empleando los SIMM de 72 contactos, una CPU de 32 bits tan sólo requerirá de un módulo por banco de memoria.

La característica principal que distingue a los módulos SIMM se encuentra en los terminales de conexión a la placa base. Los terminales de una cara del módulo se hallan conectados a sus homónimos en la cara opuesta.

.

Módulos DIMM

La Figura 3 muestra el aspecto de un módulo DIMM. Como se puede apreciar, es muy similar al de los módulos SIMM. Sin embargo, hay una diferencia importante: los terminales de conexión de una cara están eléctricamente aislados de sus equivalentes en la cara opuesta. Esto permite disponer de mayor número de terminales y, por tanto, de mayor capacidad de transferencia de datos. De hecho, los módulos DIMM se suelen emplear con las CPU capaces de procesar 64 bits de forma simultánea (como es el caso de los procesadores Pentium). Estos módulos presentan 168 contactos, y las distintas configuraciones proporcionan palabras de 32 ó 64 bits, y presentan capacidades de almacenamiento de 32, 64, 128 MB...

Figura 3. Módulo DIMM de 168 contactos.

Los módulos SODIMM (*Small Outline DIMM*, DIMM de contorno pequeño) se encuentran con frecuencia en los ordenadores portátiles. Los módulos SO-DIMM de 72 contactos (Figura 4) se asemejan a los SIMM de 72 contactos, pero con dimensiones más reducidas. La reducción de tamaño se hace posible gracias al aislamiento entre caras que ofrecen los DIMM, tal y como se ha comentado antes. También existen módulos SODIMM de 144 contactos (Figura 5), en los que llama la atención la pequeña diferencia de tamaño respecto a los módulos de 72 contactos.

Figura 4. Módulo SODIMM de 72 contactos

Figura 5- módulo SODIMM de 144 contactos

Integridad de los datos

La memoria no deja de ser un circuito electrónico real, y por tanto está expuesta a efectos que pueden producir errores en su contenido. En otras palabras, tras escribir una palabra en una posición de memoria, es perfectamente posible que algún bit cambie de estado durante el tiempo que permanezca almacenada. Si se accede de nuevo a la memoria para leer dicha palabra, se recuperará información errónea, y esto puede acarrear todo tipo de consecuencias. Para ello se suelen emplear dos soluciones: la paridad y la técnica ECC *(Error Correction Code).* El elemento que implementa estos métodos se encuentra en el interior del PC, y recibe el nombre de controlador de memoria.

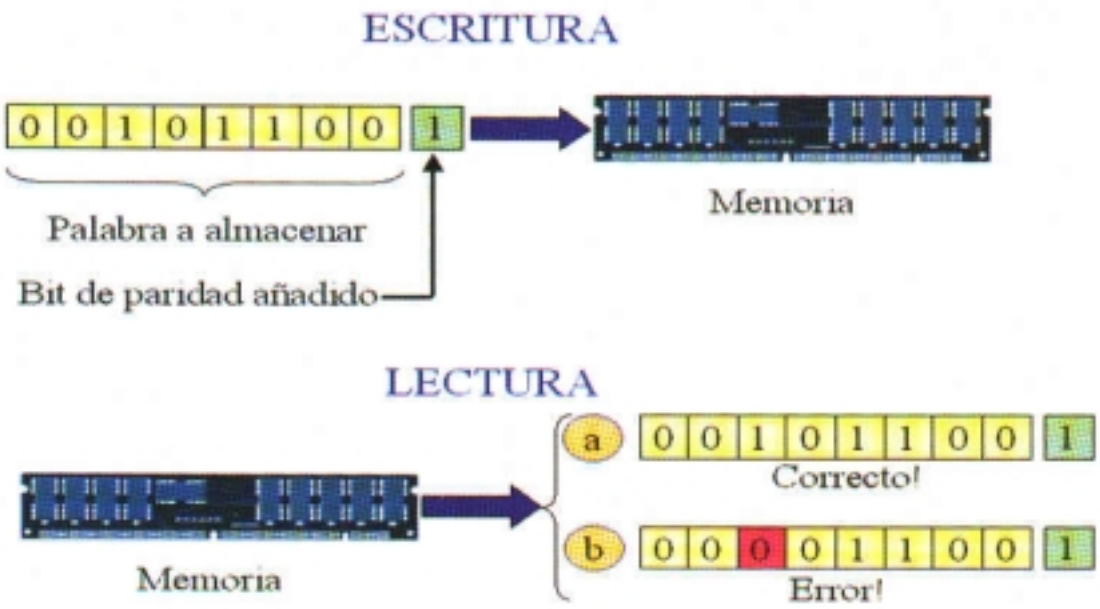


Figura 6. Detección de errores mediante paridad par. La palabra a almacenar tiene un número impar de unos. El bit de paridad se fija a 1 para forzar que la cantidad de unos sea par, antes de almacenar la palabra en memoria. Al leer dicho byte, se contará el número total de unos (incluyendo el bit de paridad). Se detectará un error si dicha cantidad no es par (en la figura se ha modificado el estado del bit resaltado en rojo)

La paridad consiste en añadir un bit adicional a cada palabra, que hace que el número de unos sea par o impar (según se emplee paridad par o impar). Si al leer información de la memoria el bit de paridad no está de acuerdo con el número de unos, se habrá detectado un error (Figura 6).

El sistema ECC añade un conjunto de bits a cada palabra a almacenar. La ventaja es que permite detectar errores en varios bits, y además es capaz de corregir dichos errores.

Estas técnicas implican añadir bits adicionales, y por tanto tendrán impacto en la cantidad de memoria incorporada en cada módulo.

DRAM y sus tipos

Las memorias RAM más empleadas son de tipo DRAM (*Dynamic RAM)*. En éstas, se debe proporcionar corriente eléctrica a las celdas -de forma periódica- para que no pierdan la información, proceso que recibe el nombre de "refres-

co". En el otro extremo se hallan las RAM estáticas o SRAM, que no requieren de dicho proceso. Las memorias DRAM son más densas, más baratas y consumen menos energía que las memorias SRAM, pero son más lentas y requieren de un circuito de refresco. A continuación se describen brevemente las tecnologías de DRAM existentes en la actualidad:

- *EDO (Extended Data Output).* Incorpora varias innovaciones, que permiten acelerar hasta un 15% la velocidad de acceso a memoria.
- *SDRAM (DRAM Síncrona)*. Emplea una señal de reloj, sincronizada con la señal de reloj de la CPU, para coordinar las transferencias. Consigue superar la velocidad de la memoria EDO en un 25%.
- *DDR (Double Data Rate) o SDRAM II*. Segunda generación de las memorias SDRAM. Aprovecha tanto el estado alto como el bajo de la señal de reloj para sincronizar las transferencias. Consigue duplicar la velocidad de transferencia sin modificar la frecuencia de reloj (es decir, con la misma CPU).
- *RDRAM (Rambus DRAM).* Esta tecnología consigue multiplicar por 10 la velocidad de transferencia de una DRAM estándar, gracias a una tecnología exclusiva denominada RSL *(Rambus Signaling Logic)*.
- *SLDRAM (synctink).* Tecnología desarrollada por un consorcio de doce compañías. Es una extensión de la SDRAM, que permite ampliar el número de bancos de 4 a 16.

Por supuesto, para aprovechar una de estas tecnologías, el PC debe estar preparado para ello. Valgan como ejemplo los procesadores Pentium 4, que por medio de un acuerdo entre Rambus e Intel, vienen preparados para emplear RDRAM.

CAPÍTULO 6

LA MEMORIA CACHÉ

Si hay algo que limita la velocidad de las transferencias entre la RAM y la CPU es sin duda la primera (no se debe olvidar que el subsistema más rápido de un PC es la CPU). Por tanto, para conseguir "acelerar" un PC no es solución sencilla, rápida ni rentable centrarse en diseñar procesadores más potentes. Aunque emplear tecnologías RAM más rápidas mejora el rendimiento, la solución óptima consiste en agregar un nuevo elemento al PC: la memoria caché. Este capítulo introduce el importante concepto de caché que todo PC actual emplea (y no sólo con la memoria principal, sino también con otros importantes subsistemas).

¿Qué es una caché?

Imagine dos sistemas de memoria A y B, entre los cuales se transfiere información. Suponga que el sistema A es más rápido y presenta menor capacidad de almacenamiento que B (situación típica en un PC). Esto se traduce en que A debe funcionar a menor velocidad de la que es capaz de ofrecer, siempre que se comunique con B. Se puede mejorar la velocidad de transferencia introduciendo un nuevo sistema de memoria C entre A y B, al que se denomina caché. La caché debe presentar una capacidad de almacenamiento mayor que la de A y menor que la de B. Además, será más lenta que A, pero más rápida que B. En otras palabras, sus características son un término medio entre los sistemas A y B. La aceleración de las transferencias se basa en almacenar la información intercambiada últimamente entre A y B, puesto que con gran probabilidad será la más empleada en las próximas transferencias. La aplicación de sistemas de caché recibe el nombre de caching.

Esto ya se introdujo al colocar la memoria RAM entre la CPU (rápida y con poca capacidad) y otros dispositivos lentos (y de gran capacidad), como es el caso del disco duro.

Ya que esta definición es quizá un tanto confusa, en el siguiente apartado presentamos más detalles acerca del funcionamiento de una caché, que permiten entender el concepto con mayor claridad. Tomaremos como referencia un tipo particular de caché: la caché de memoria. Ésta se introduce entre la RAM y la CPU. De esta forma, se consigue incrementar notablemente la velocidad con que la CPU accede a la memoria principal. Por supuesto, la memoria caché será más rápida que la RAM, y dispondrá de menor capacidad de almacenamiento. Veamos cómo funciona.

¿Cómo funciona una caché de memoria?

En realidad, el funcionamiento de una caché sigue un principio parecido al que formulamos para la memoria principal. En aquel caso, las instrucciones y datos

se cargaban en la RAM (desde dispositivos lentos), donde la CPU podría acceder a mayor velocidad.

Una caché de memoria se carga (desde la RAM) con los datos y/o instrucciones que ha buscado la CPU en las últimas operaciones. La CPU buscará siempre primero la información en la caché, y la encontrará allí la mayoría de las veces, con lo que el acceso será muy rápido. Si, desgraciadamente, no se encuentra la información en la caché, se perderá un tiempo extra en acudir a la RAM y copiar dicha información en la caché para que esté disponible (ver Figura 1). Como estos fallos ocurren con una frecuencia relativamente baja, el rendimiento mejora considerablemente, ya que la CPU accederá más veces a la caché que a la RAM. Lo comentado para la búsqueda en memoria funciona de forma análoga para la escritura: la CPU escribe en la caché en lugar de la RAM (más adelante hablaremos de dicho proceso con mayor detalle).

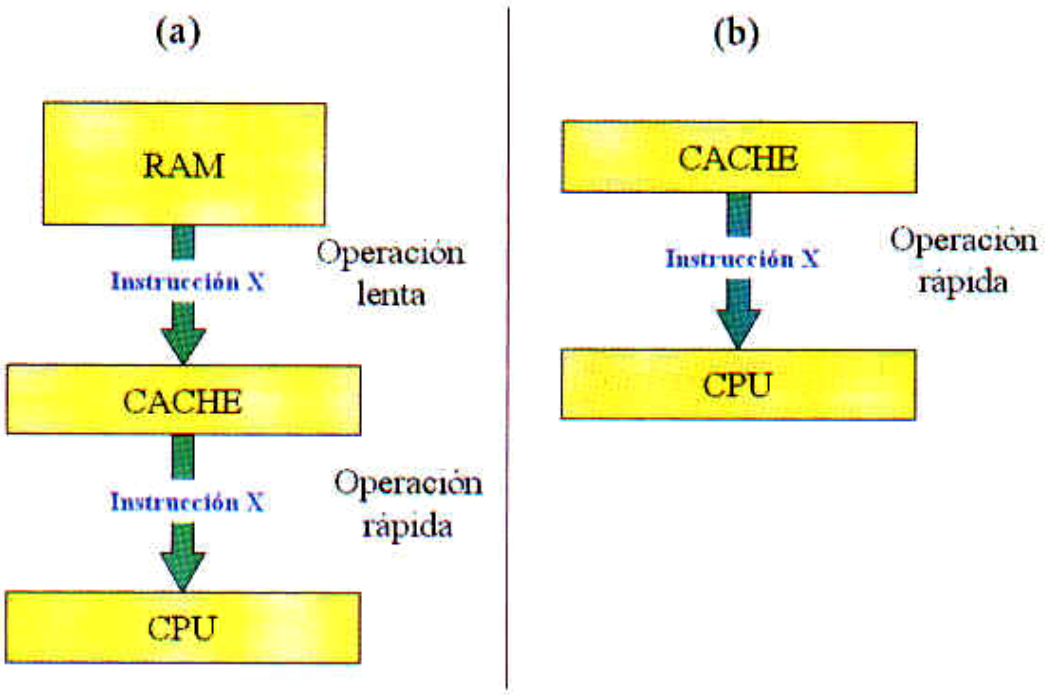


Figura 1. Operación de lectura empleando cahé. Si la CPU desea buscar la instrucción X (situada en cierta posición de la RAM), intentará hallar primero dicha instrucción en la caché. Si la encuentra allí, el acceso será rápido (caso B). Si no la encuentra, se añade un retraso extra al buscar la instrucción en la RAM y copiarla en la caché.

No es de extrañar que esta técnica (caching) funcione. La mayoría de programas tiene un alto contenido en bucles, lo que implica un uso repetido de instrucciones (e incluso datos). Y esto implica una elevada probabilidad de acierto al buscar la información en la caché.

Una forma interesante de ilustrar el funcionamiento de la caché consiste en imaginar un videoclub, equipado con un mostrador y una habitación capaz de almacenar cientos de vídeos. Ante la petición de cada cliente, el dependiente deberá acudir hasta el almacén, buscar la película solicitada, volver al mostrador y entregar la cinta al cliente. Ante la devolución de una cinta, el dependiente debe caminar hacia el almacén y guardar dicha cinta en el lugar apropiado.

Realmente esa forma de trabajo no es nada eficiente, ya que implica demasiados desplazamientos y, por tanto, la atención al cliente será realmente lenta.

Suponga ahora que el dependiente dispone de un pequeño archivador de 20 vídeos sobre el mostrador. Cuando un cliente devuelve una cinta, el dependiente coloca la cinta directamente en el archivador, en lugar de caminar hacia el almacén. Si se va repitiendo dicho proceso, el dependiente dispondrá continuamente de las veinte últimas películas devueltas en el archivador. Cuando se acerque un cliente y pida una película, el dependiente la buscará primero en el archivador, y sólo si no la encuentra allí se desplazará hacia el almacén. Este método funciona, sobre todo porque la mayor parte de las películas devueltas serán las de estreno, que al mismo tiempo son las más solicitadas.

Otros tipos de caché

Aunque hasta ahora hemos hablado de la caché con respecto a la memoria RAM, en un PC existen muchos otros sistemas de caché.

Sin ir más lejos, las unidades de almacenamiento (discos duros, discos flexibles, etc.) y otros muchos periféricos utilizan la memoria RAM como sistema de caché. En efecto, una zona de la RAM contiene la información que se ha buscado últimamente en dichos dispositivos, de forma que basta con acceder a la RAM para recuperarla. La escritura funciona de forma análoga: se escribe información directamente en la RAM, y ésta se vuelca a las unidades asociadas cuando es oportuno. Evidentemente, el rendimiento mejora de forma notable. Incluso es posible emplear el disco duro como caché de cara a dispositivos aún más lentos (como son las unidades CD-ROM). Estos sistemas de caché suelen estar gobernados mediante software, que se suele integrar en el sistema operativo.

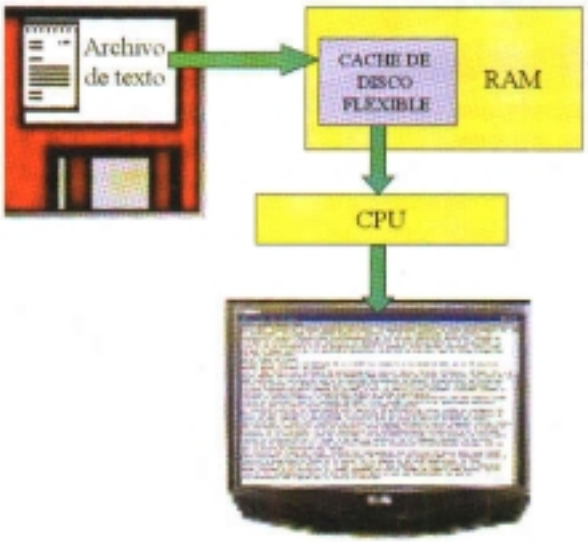


Figura 2. Primer acceso a un archivo de texto, almacenado en un disco flexible. La información no se halla en la caché del disco flexible (porción de RAM). Por ello, es necesario copiarla primero del disco a la RAM, y de ahí el retardo inicial.

Es sencillo realizar un experimento para apreciar la presencia de estos tipos de caché en cualquier PC. Basta con que introduzca un disco flexible en la unidad

correspondiente con un archivo de texto de unos 300 kB. A continuación, inicie el bloc de notas y abra el fichero de texto. Apreciará que el indicador luminoso de acceso a la unidad permanece varios segundos activo, hasta que finalmente aparece el texto en pantalla. Acto seguido, cierre el bloc de notas y repita el proceso. En esta ocasión, el texto aparecerá casi instantáneamente en pantalla y el indicador luminoso no se encenderá. ¿Qué ha ocurrido? En la primera ocasión se ha acudido al disco para copiar la información en la RAM (ver Figura 2), y de ahí la tardanza. En el segundo acceso, el sistema operativo ha buscado directamente en la caché asociada al disco flexible y ha encontrado la información buscada, por lo que el acceso es mucho más rápido (ver Figura 3).

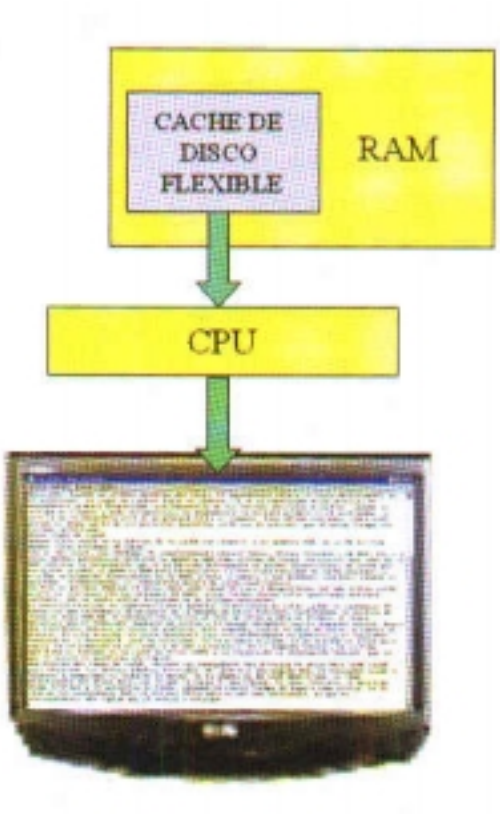


Figura 3. Segundo acceso al mismo archivo de texto. En esta ocasión, se encuentra la información en la caché asociada al disco flexible, por lo que el retado es mucho menor.

Aún existen más tipos de caché. Incluso los navegadores Web utilizan el disco duro como caché para almacenar las últimas páginas visitadas. Al solicitar una página Web, el navegador acude a Internet y comprueba la fecha de la misma. Si la página no ha sido modificada, se toma directamente del disco duro, con lo que la carga es muy rápida. En caso contrario se descarga desde Internet y se actualiza la caché, pagando un cierto tiempo de espera como precio. En el caso de los navegadores Web, el uso del disco duro es más que suficiente, ya que es extremadamente más rápido que el acceso a Internet.

Niveles de caché

Tal y como acabamos de mostrar, un PC incorpora varios tipos de caché. Pero, ¿de qué forma están organizados? Usualmente, los diferentes sistemas de ca-

ché se organizan por niveles, formando una jerarquía. En general se cumple que, a mayor cercanía a la CPU, se presenta mayor velocidad de acceso y menor capacidad de almacenamiento (ver Figura 4).

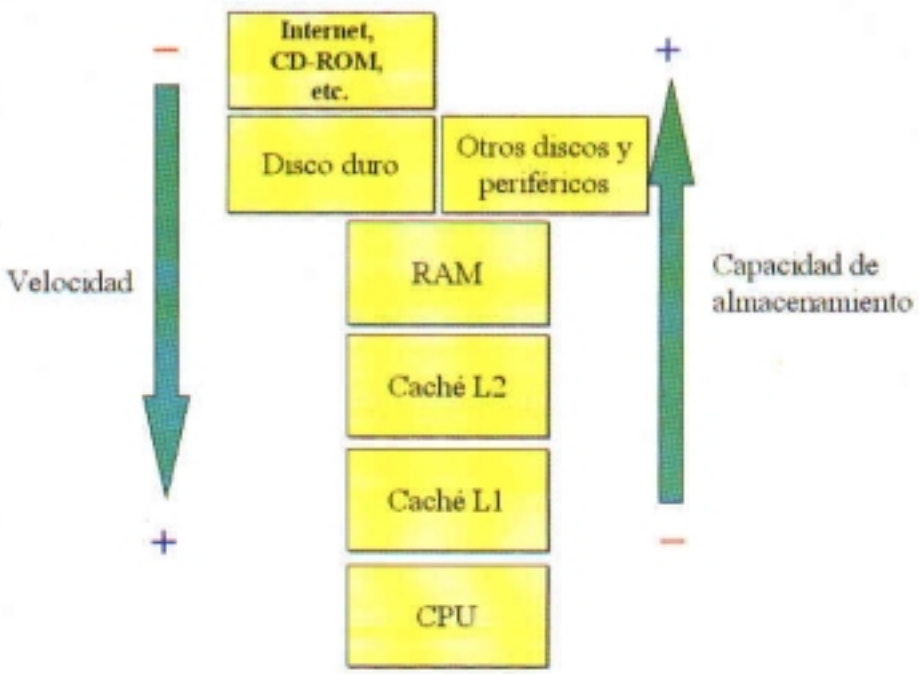


Figura 4. Niveles de caché. Nótese que incluso el disco duro se puede emplear como caché para otros dispositivos (como los CD-ROM) y aplicaciones (como los navegadores Web).

Para empezar, la caché de memoria se suele desglosar en dos niveles. En el nivel más cercano a la CPU se encuentra la caché L1 (level 1 o nivel 1). Ésta se halla integrada en el mismo chip que la CPU, con lo que el acceso se produce a la velocidad de trabajo del procesador (la máxima velocidad). Por supuesto, la caché L1 presenta un tamaño muy reducido (de 4 a 16 kB).

A continuación aparece la caché de nivel 2 o L2. Inicialmente, se instalaba en la placa base, en el exterior de la CPU. Los procesadores actuales la integran en el propio chip. Como era de esperar, tiene mayor capacidad que la caché L1 (de 128 a 512 kB) pero es una memoria más lenta. Por ejemplo, en el procesador Pentium II la velocidad de acceso era la mitad respecto a la caché L 1.

El siguiente nivel lo constituye la memoria RAM, que ya tratamos en la anterior entrega. Como ya hemos comentado, la RAM suele hacer de caché para los dispositivos de almacenamiento y otros tipos de periféricos. El nivel más alto lo ocuparían las caché en disco duro, como son las utilizadas por los navegadores Web.

La búsqueda de información comienza por la caché L1, y se va subiendo nivel a nivel en caso de no encontrar lo que se busca en el nivel actual. Por supuesto, cuantas más capas haya que ascender, mayor retardo se pagará. Pero, a mayor cercanía a la CPU, la probabilidad de encontrar lo que se busca es ma-

yor. Esta forma de trabajo resulta una excelente relación de compromiso entre diversos factores, y consigue mejorar el rendimiento del ordenador de forma notable.

Estructura y funcionamiento interno de una caché de memoria

No es el objetivo de este trabajo presentar con detalle la arquitectura interna de una caché L2, pero es conveniente proporcionar un conocimiento general sobre lo que ocurre en su interior.

Al igual que ocurría con la RAM, es apropiado pensar en la caché como un arreglo de tipo tabla. En este caso, cada celda almacena un octeto o byte. No es raro que una caché de 512 kB se distribuya en 16.384 filas (16 kB) y 32 columnas (32 bytes).

La entidad básica de almacenamiento la conforman las filas, a las que se llama también "líneas de caché". En el ejemplo anterior se dispone de 16.384 líneas de caché, de 32 bytes cada una. Nunca se toma un byte de la RAM y se escribe en una celda de la caché. Por el contrario, en cada movimiento siempre se copia información de la RAM suficiente para cubrir una línea de caché (en el ejemplo, siempre se mueven 32 bytes). En el caso de la escritura, el funcionamiento es totalmente análogo.

Toda caché incorpora, además, un espacio de almacenamiento llamado Tag RAM, que indica a qué porción de la RAM se halla asociada cada línea de caché. En otras palabras, la Tag RAM permite traducir una dirección de RAM en una línea de caché concreta.

Ya que la RAM tiene mayor capacidad que la caché, ¿cómo se reparte la RAM entre las líneas de caché disponibles? Existen tres tipos de caché según la técnica empleada:

Caché de mapeo directo. Se divide la RAM en porciones de igual tamaño, tantas como líneas de caché existan. Cada línea de caché es un recurso a compartir por las direcciones de memoria de una porción diferente. Por ejemplo, si se dispone de una RAM de 64 MB y la caché de 512 kB presentada anteriormente, cada línea podrá almacenar 32 de las 4.096 direcciones que contiene la porción de RAM asociada (64 MB/16.384 líneas = 4.096 bytes / línea). Esta técnica permite una búsqueda muy rápida, ya que cada posición de RAM sólo puede estar en una determinada línea. Sin embargo, la probabilidad de encontrar la información buscada es mínima. Imagine dos instrucciones A y B, que se corresponden con la misma línea de caché (esto es, pertenecen a una misma porción de RAM). Suponga que la CPU necesita ejecutar una secuencia alternada A, B, A, B, etc. En ese caso, se tendrá que acceder a la RAM para copiar A y luego para copiar B (y reemplazar a la instrucción A en la caché), y así hasta terminar la secuencia. Sin duda, el porcentaje de acierto es nulo en dicha situación.

Caché completamente asociativa. Cada línea de caché se puede llenar con cualquier grupo de posiciones de la memoria RAM. En este caso, el porcentaje de acierto es máximo, y el ejemplo anterior no produciría problemas. En cambio, el tiempo de acceso es muy elevado, puesto que una posición de RAM

puede estar en cualquier línea de caché (esto es lento, incluso empleando algoritmos de búsqueda avanzados).

Caché asociativa por conjuntos de N líneas. La caché se divide en conjuntos de N líneas. A cada conjunto se le asocia un grupo de posiciones de RAM. Dentro del conjunto asignado, una posición de RAM puede ir a parar a cualquiera de las N líneas que lo forman. En otras palabras, dentro de cada conjunto la caché es totalmente asociativa. Esta situación es la más equilibrada, puesto que se trata de un compromiso entre las técnicas anteriores. Si se hace N=1, se tiene una caché de mapeo directo. Si N es igual al número de líneas de la caché, se tiene una caché completamente asociativa. Si se escoge un valor de N apropiado, se alcanzará la solución óptima.

Normalmente, la caché L2 es de mapeo directo, mientras que la caché Ll es asociativa por conjuntos de N líneas.

Políticas de escritura

El proceso de escritura en caché es muy simple: en lugar de escribir la información en la RAM, se escribe directamente en la caché. El detalle a resol-ver es: ¿cuándo se traslada la información de la caché a la RAM? Hay dos políticas de escritura fundamentales para resolver dicho problema:

Write-Back. La información se escribe directamente en la caché, sin actualizar la RAM. Cuando una posición de la caché debe ser utilizada por otra posición de RAM diferente, su contenido actual se traslada a la RAM, asegurando la coherencia entre ambas memorias.

Write-Through. Cada vez que se escribe en una línea de caché, se actualiza la RAM. Esta técnica conlleva un acceso continuo a la RAM, por lo que el rendimiento es pobre.

Las caché Write-Back proporcionan el mayor rendimiento, pero conlleva un riesgo de integridad. Por ejemplo, en un momento dado, el contenido de la RAM y la caché L2 pueden ser diferentes.

Con la memoria RAM esto no tiene gran importancia. Pero en casos como la caché asociada al disco duro (espacio en RAM), y ante un fallo de la alimentación eléctrica, esto puede implicar mantener en disco una información no actualizada. Por ello, las cachés de disco suelen evitar la técnica Write-Back.

CAPÍTULO 7

LAS INTERFACES IDE Y SCSI

Aunque los periféricos estándares que incorpora un PC son apropiados para algunas aplicaciones, existen otras muchas en las que es necesario instalar nuevo hardware. Las ranuras de expansión (ISA o PCI) permiten conectar nuevos periféricos pero, lamentablemente, su número no es muy generoso. Esto constituye una limitación, ya que en muchas ocasiones es interesante instalar más dispositivos que los permitidos por dichas ranuras. Por ello, el PC ofrece otras vías de expansión, como son los puertos serie y paralelo. Estos puertos aumentan la capacidad de expansión, pero se caracterizan por su lentitud, lo cual los hace inapropiados para conectar ciertos tipos de dispositivos.

Todo esto ha hecho que nazcan nuevas interfaces para la conexión de hardware en el PC. En este artículo se abordarán dos de las interfaces más conocidas: IDE (ampliamente utilizada para la conexión de discos duros y unidades CD-ROM) y SCSI (caracterizada por su gran rendimiento y capacidad de expansión).

La interfaz IDE

El término IDE (*Integrated Drive Electronics*) procede del año 1986, cuando las firmas Compaq Corporation, Western Digital y Control Data Corporation trabajaban juntas en un proyecto común. Se trataba de integrar un chip controlador fabricado por Western Digital en una unidad de disco duro. En 1988, se formó un grupo industrial denominado CAM (*Common Access Method* o método de acceso común), el cual desarrolló un estándar que cubría la integración de dispositivos controladores en unidades de almacenamiento, y su conexión al PC. Dicho estándar fue aprobado en 1991, bajo el nombre de ATA (*AT Attachment*).

Aunque hoy en día se utiliza el término IDE para referirse a ATA (y en el presente capítulo así se hará), es importante remarcar la diferencia que existe realmente entre ambos acrónimos. Mientras que IDE se refiere a las unidades de almacenamiento que integran el circuito controlador asociado, ATA hace referencia a la interfaz para interconectar los dispositivos IDE y el PC.

Tal y como acabamos de introducir, en una unidad de almacenamiento IDE el dispositivo controlador correspondiente se encuentra integrado en la propia unidad. Esto hace que sean necesarios menos componentes, y que la integración entre unidad y controlador sea óptima, y realizada por el fabricante. Como se puede intuir, esto proporciona muchas ventajas.

En primer lugar, la conexión al bus del sistema es realmente simple. Dicha conexión se suele realizar de forma directa, mediante conectores soldados sobre la placa base. Esto evita utilizar ranuras de expansión, dejándolas libres para otros dispositivos.

Además, el coste de producción de una placa base con conectores IDE es menor que el que implica disponer de una tarjeta controladora.

Otro factor importante es la reducción del número de cables necesarios, ya que la unión entre dispositivo y controlador ya viene implementada en el propio dispositivo. El controlador -al estar integrado- se halla conectado al dispositivo mediante conexiones de pequeña longitud, consiguiendo que la resistencia a interferencias sea óptima, y en general mejores prestaciones.

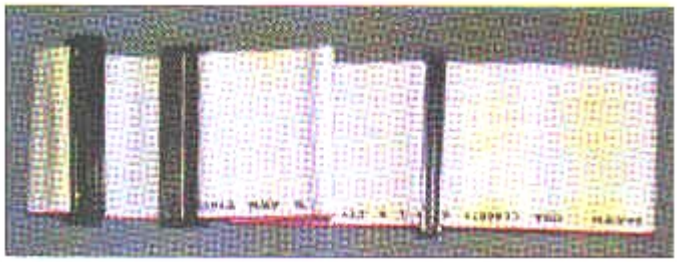


Figura 1. Cable IDE

Por otro lado, el fabricante no se debe preocupar por respetar ninguna interfaz estándar entre el controlador y el dispositivo, detalle que flexibiliza el diseño y permite, así, obtener mejores productos. En otras palabras, cada unidad y su controlador forman un producto independiente.

Todos estos detalles justifican que la mayoría de placas base actuales incorporen conectores IDE.

Conectores y cables IDE

Un cable IDE estándar presenta tres conectores: uno de ellos se une a un conector IDE de la placa base, y los dos restantes (cercanos entre ellos) permiten conectar dos dispositivos IDE (ver Figura 1). Hay que anotar que existen otras posibles configuraciones, pero la expuesta aquí es la más común. El cable es de tipo cinta y plano, con 40 hilos colocados en paralelo y aislados entre sí. El hilo correspondiente a una de las extremidades del cable se halla coloreado en rojo. Dicha parte del cable se conecta al pin número 1 del conector de la placa base, y también de los dispositivos. El cable no debe superar los 45 centímetros de longitud.

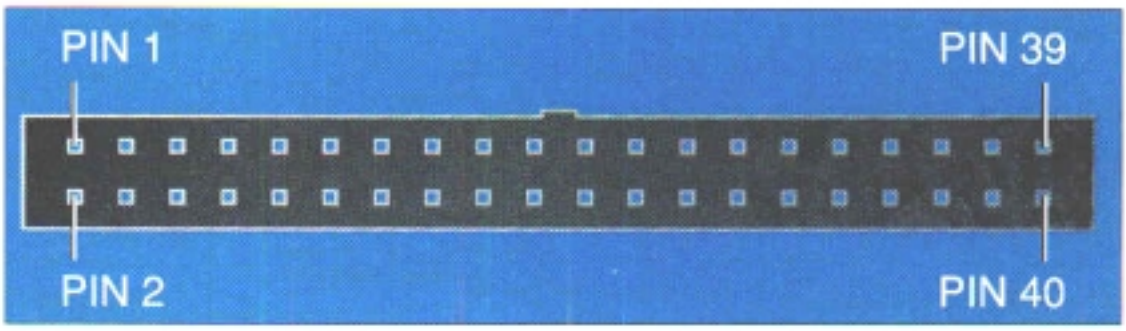


Figura 2. Conector IDE (40 pines)

Cada conector IDE presenta 40 puntos de conexión (normalmente denominados "pines"). El aspecto de un conector IDE se presenta en la Figura 2.

Si ya se han ocupado todos los conectores IDE disponibles, y se desea instalar más dispositivos, existe una posible solución: instalar una tarjeta controladora IDE en una ranura de expansión del PC. Ésta proporciona dos conectores IDE adicionales, lo que permite instalar 4 dispositivos más.

Las últimas versiones del bus IDE, trabajando a 66 MB/s o más (ATA 66, ATA 100...) precisan de un cable especial con 80 hilos en lugar de cuarenta, aunque mantiene el conector de 40 contactos. Los cuarenta cables extra están conectados a masa y permiten asegurar la integridad de los datos a altas velocidades.

Configuración de Jumpers

Muchos dispositivos IDE soportan tres tipos de configuraciones: dispositivo simple, maestro o esclavo. Estos modos se suelen seleccionar mediante una pequeña serie de conmutadores o jumpers, que suelen aparecer en la parte trasera del dispositivo. El modo simple indica que la unidad está sola en el sistema, y por tanto responde a todos los comandos IDE recibidos. Cuando hay dos unidades en el mismo cable IDE, una se configura como maestro y la otra como esclavo. La unidad maestra responderá únicamente a los comandos IDE maestro, y la unidad esclava hará lo mismo con los comandos esclavo.

La interfaz EIDE

La interfaz IDE original gozaba de unas prestaciones que debían ser mejoradas. Como solución, apareció la interfaz ATA-2 , conocida como EIDE (*Enhanced IDE*). Fue propuesta por las firmas Western Digital y Seagate Technologies, y es la versión de IDE que se encuentra hoy en día en el mundo del PC. Las principales mejoras respecto a IDE son las siguientes:

Mayor capacidad de almacenamiento. Los avances en las BIOS permitieron trabajar con unidades de más de 504 MB.

Mayor número de discos duros. Es posible incorporar más de dos unidades de disco.

Mayor velocidad. La técnica de entrada/ salida programada (PIO, *Programmed Input/Output*) permite seleccionar entre varios modos de trabajo, que consiguen alcanzar relaciones de transferencia de hasta 16,6 MB por segundo.

ATAPI (ATA Paeket Interface). Permite conectar otros tipos de unidades de almacenamiento a un conector IDE (por ejemplo, unidades de CD-ROM).

Soporte DMA. La interfaz ATA-2 soporta DMA (*Direct Memory Access*, acceso directo a memoria). Como su nombre indica, DMA permite intercambiar información entre las unidades de almacenamiento y la memoria, sin pasar por la CPU. La CPU no debe preocuparse de las transferencias (quedando libre para otras tareas) y, por tanto, la velocidad de transferencia es claramente mayor y el rendimiento del PC mejora significativamente.

La interfaz SCSI

Además de IDE, hay otra interfaz que goza de una enorme aceptación, denominada SCSI (*Small Computer System Interface*). Para hacer más cómoda su pronunciación, dicho acrónimo se suele leer como "scuzzy" en inglés, lo que en castellano se pronunciaría "escasi".

La interfaz SCSI permite al PC intercambiar datos con todo tipo de dispositivos: discos duros, CD-ROM, impresoras, etc. Algunos PC soportan SCSI en la propia placa base, pero no se trata de la opción más usual. Normalmente, es necesario instalar una tarjeta adaptadora SCSI en una de las ranuras de expansión del sistema, que es la que permite la conexión de los dispositivos (interna o externamente).

Dicha tarjeta es fácil de encontrar y se instala de forma sencilla.

Una de las principales ventajas de SCSI es el gran número de dispositivos que puede controlar. Mientras que IDE sólo soporta dos unidades y EIDE llega hasta cuatro, SCSI permite la conexión de hasta 8 dispositivos (incluyendo la tarjeta controladora SCSI), utilizando tan sólo una ranura de expansión. Además, la velocidad de transferencia es superior a la que caracteriza a la interfaz IDE.

Si se desea aumentar la capacidad de expansión, se puede instalar una segúnda tarjeta controladora SCSI, lo que permite conectar 7 periféricos más. Mejor aún, existen tarjetas controladoras que soportan 15 periféricos, consumiendo tan sólo una ranura de expansión.

Por supuesto, es posible instalar discos IDE y SCSI simultáneamente en un PC. La unidad IDE seguirá siendo el disco de arranque y los dispositivos SCSI proporcionarán capacidad de almacenamiento adicional.

También es interesante señalar que, en caso de no disponer de ranuras de expansión libres, existen adaptadores que permiten conectar dispositivos SCSI al puerto paralelo. Los dispositivos trabajarán a una velocidad considerable-mente menor, pero esta solución puede resultar interesante en algunos casos.

Versiones de SCSI

Aunque SCSI se considera un único estándar, se presenta en diferentes variantes: SCSI, SCSI-2, SCSI-3, Fast SCSI, Fast Wide SCSI, Ultra SCSI, Wide Ultra SCSI, Ultra2 SCSI, Wide Ultra2 SCSI, Ultra 3 SCSI, Ultra4 SCSI, etc.

Las diferencias principales radican en las longitudes de cable permitidas, las velocidades de transferencia alcanzadas y el número de dispositivos soportados. El modificador "Wide" indica que se trabaja con 16 bits en lugar de ocho, permitiendo la conexión de 15 periféricos. La variante más actual se encuentra en discos duros muy rápidos y se denomina Ultra320, alcanzando 320 MB por segundo.

Otra diferencia importante es el tipo de señales soportadas, que afecta directamente a la longitud de cable permitida. Existen tres formas de transportar las señales eléctricas por los cables: single-ended, LVD (*LoWloltage Differential*) y HVD (*High Uoltage Differential*). El modo single-ended utiliza una línea para transportar los datos y la otra línea se conecta a masa. Las variantes por deba-

jo de Ultra2 SCSI permiten longitudes de cable de 3 ó 6 metros en dicho modo (según variante). Para las variantes Ultra2 SCSI y superiores, el modo single-ended no se encuentra definido.

Los modos diferenciales (LVD y HVD) usan dos líneas de datos, y la información se transmite como la diferencia de tensión entre ambas líneas. El modo diferencial es más resistente a las señales ruidosas, por lo que se permitirán mayores longitudes de cable que en modo single-ended. La ventaja del modo LVD es que requiere menor consumo de potencia en el envío de señales, y resulta más barato que el trabajo en modo HVD. La ventaja del modo HVD es que tra-baja con mayores voltajes, y por tanto permite emplear cables más largos.

LVD soporta longitudes de cable de 12 metros, mientras que HVD admite 25 metros. Por ello, los modos diferenciales se suelen emplear cuando los dispositivos están esparcidos a través de una habitación. En general, los dispositivos single-ended son más baratos que los diferenciales.

La compatibilidad entre las diferentes versiones de SCSI está asegurada en ambos sentidos. Si se usa una tarjeta controladora SCSI antigua con un dispositivo SCSI más moderno, este último funcionará a la máxima velocidad que permita la tarjeta (por tanto, quedará ralentizado). En el caso opuesto (tarjeta moderna y dispositivo antiguo), el dispositivo SCSI funcionará a la máxima velocidad que le sea posible (desaprovechando la velocidad que es capaz de dar la tarjeta controladora).

Conectores SCSI

Los dispositivos y tarjetas controladoras SCSI pueden presentar diferentes tipos de conectores.

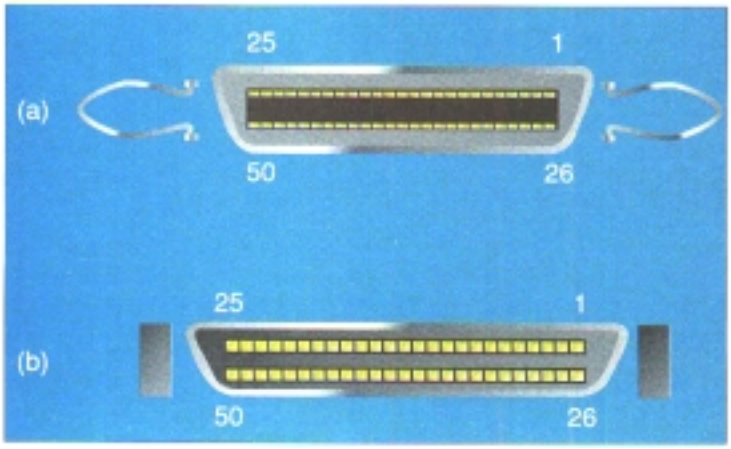


Figura 3. Conectores SCSI de 50 contactos

El conector tipo Centronics de 50 contactos (Figura 3-a) se utiliza con las variantes SCSI, Fast SCSI y Ultra SCSI a 8 bits. Este conector se conoce también bajo el nombre de conector SCSI-I, y está presente en multitud de discos duros, escáneres y grabadoras de CD. Según indica la especificación, la denominación del cable correspondiente es *Alt2, A-cable conector*.

Existe otro conector SCSI de 50 contactos de alta densidad, también conocido como conector SCSI-2 (Figura 3-b). La especificación del cable es *Altl, A-cable conector*. Como se puede apreciar, las sujeciones del conector son de tipo *latch* (enclavamiento), en lugar de tornillos.

También con 50 contactos, se encuentra el conector DB-50 (Figura 4-a). Este conector no se encuentra reconocido por la especificación SCSI, y se empleaba en antiguas computadoras Sun.

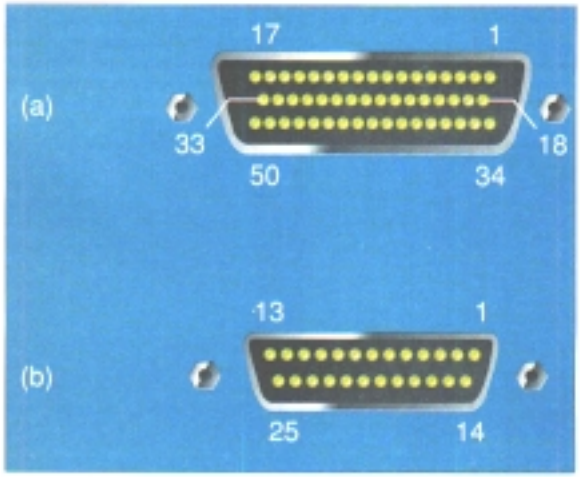


Figura 4. Conectores DB50 y DB25

De forma similar, el conector DB2S (Figura 4-b) -con 25 contactos- se empleaba en dispositivos orientados a computadoras Apple y algunas estaciones de trabajo Sun. Este conector tampoco está reconocido por la especificación SCSI, y además admitía una menor longitud de los cables y velocidad de transferencia.

La Figura 5 muestra el conector SCSI de 68 contactos de alta densidad, conocido como "conector SCSI-3". Este conector se emplea con dispositivos SCSI rápidos, tipo Fast Wide SCSI y Wide Ultra SCSI. La especificación del cable es *Alt3, P-cable conector*.

También existe una versión de conector de 68 contactos de muy alta densidad, conocido como VHD-CI. Éste presenta unas dimensiones muy reducidas, y permite colocar hasta 4 conectores sobre una tarjeta ISA o PCI. En este caso, el cable se denomina *Alt4, P-cable conector*.

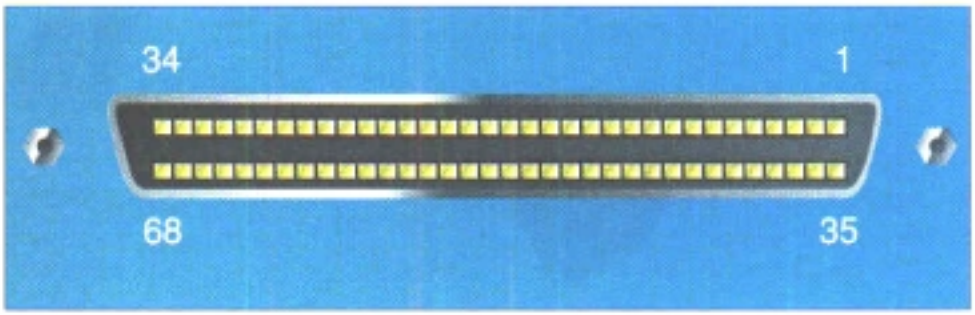


Figura 5. Conector SCSI de 68 contactos de alta densidad

El bus SCSI

Los dispositivos SCSI se conectan a la tarjeta controladora en forma de cadena, definiendo un bus que opera de forma independiente al resto del PC. En efecto y al contrario que en la interfaz IDE- el bus SCSI permite que los dispositivos intercambien información sin necesidad de consumir tiempo de procesamiento de la CPU. Esto último explica la superioridad de SCSI sobre IDE en términos de velocidad.

Cada dispositivo SCSI dispone de dos conectores, de forma que uno de ellos se conecta al dispositivo anterior en la cadena, y el otro se conecta al dispositivo siguiente. Uno de los extremos de la cadena se une al conector externo de la tarjeta controladora. Ésta es la configuración típica, en que la tarjeta controladora forma uno de los extremos de la cadena.

Además, la tarjeta controladora dispone de un segundo conector, destinado a la conexión de dispositivos internos. Estos formarán otra cadena -instalada en el interior del PC- y, por tanto, la tarjeta adaptadora ya no quedará en un extremo.

Una vez la cadena ha sido implementada, es necesario conectar unos elementos denominados terminadores en sus extremos. La instalación de terminadores es obligatoria, ya que se trata de impedancias que evitan comportamientos no deseados en las señales de alta frecuencia que circulan por el bus.

Los terminadores vienen incluidos al comprar dispositivos SCSI, por lo que su adquisición no acarrea problemas.

Un dispositivo SCSI suele incorporar un terminador aplicado en el conector destinado al siguiente componente de la cadena. Por tanto, sólo habrá que quitarlo si hay que añadir otro dispositivo después de él.

La tarjeta controladora suele incorporar un terminador en el conector destinado a dispositivos internos. Sólo será necesario eliminarlo si se va a extender la cadena con dispositivos internos.

En cualquier caso, la regla a recordar es que los terminadores deben colocarse en los extremos de la cadena (que pueden ser dispositivos SCSI o la propia tarjeta controladora).

En cuanto al direccionamiento del bus, los dispositivos incluidos en la cadena se acceden -desde el punto de vista de la BIOS- a través de un identificador o ID (que irá de 0 a 7, o bien de 0 a 15, dependiendo del número de dispositivos soportados). Sin embargo, cada dispositivo puede ser interpretado como un conjunto de hasta 8 unidades lógicas. Cada unidad lógica se identifica mediante un número, denominado LUN (*Logical Unit Number)*. La mayoría de dispositivos SCSI sólo contienen una unidad lógica (y por tanto, todos ellos recibirán un LUN igual a 0). En casos tales como las unidades que integran varios lectores de CD-ROM, a cada uno de los lectores se le asignará un LUN diferente. Esto permite acceder a cada unidad lógica de forma independiente a través de los LUN.

CAPÍTULO 8

EL PRECURSOR DEL DISCO DURO: EL DISCO FLEXIBLE.

Los discos flexibles (conocidos en inglés como *floppy disks*) están muy cerca de entrar en el museo de antigüedades del mundo del PC. Fueron los elementos estrella de la distribución software desde el nacimiento del PC hasta mediados de los años 90, cuando los CD-ROM ocuparon su lugar. De hecho, en la era antigua del PC el centro de datos por excelencia era el disco flexible (el disco duro aún no se utilizaba). Aunque actualmente la mayoría de los PC vienen equipados con una unidad de disco flexible, su uso está lejos de ser frecuente, ya que su capacidad de almacenamiento y velocidad son inapropiadas para las necesidades actuales. Sin embargo, dicho medio todavía resulta útil para tareas como el traslado de pequeños archivos entre ordenadores, almacenamiento de ficheros de tamaño reducido, y la instalación de software de pequeño tamaño, como es el caso de los controladores de hardware.

La clara desventaja de los discos flexibles respecto al resto de los medios de almacenamiento no sólo radica en su baja capacidad, sino también en su bajo rendimiento. La suma de todas estas desventajas explica la disminución del precio de las unidades de disco flexible, que han llegado a convertirse en un componente realmente económico.

En resumen, los discos flexibles no se utilizan demasiado, y tienen un futuro bastante oscuro. Teniendo esto en cuenta, no parece muy apropiado dedicarle un artículo a dicho medio de almacenamiento. Pero la realidad es bien diferente: es conveniente tratar el disco flexible, ya que muchos de los conceptos que lo rodean se aplican a los discos duros.

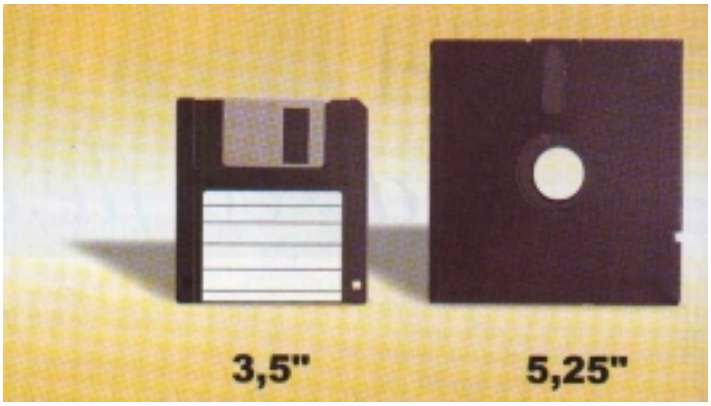


Figura 1. Discos flexibles de 3,5" (izquierda) y 5,25" (derecha). El disco de 8" tenía un aspecto parecido al de 5,25".

Formatos de disco flexible

Los primeros formatos de disco utilizados eran de 8". Sin embargo, los formatos que más se han extendido han sido el disco de 5,25" y el de 3,5" (ver Figura

1). No se debe olvidar que dichas indicaciones (en pulgadas) hacen referencia al lado del cuadrado que forma la carcasa del disco, y no al disco propiamente dicho.

En todos estos formatos, el soporte de almacenamiento es una superficie circular maleable, de ahí el nombre de disco flexible. Sobre dicha superficie se encuentra un recubrimiento de material magnético, que es el que efectivamente almacena la información. En ambos casos, el disco se introduce en el interior de una carcasa de plástico, que lo protege del exterior.

El disco de 5,25" es el más antiguo de los dos (de hecho, ya se puede dar por obsoleto), y fue introducido por la firma Shugart. El auténtico disco se albergaba en el interior de una carcasa de plástico fina y frágil. En su centro aparecía un gran agujero, utilizado por la unidad para hacer girar el disco. También se podía apreciar una ventana en la carcasa, que dejaba ver el disco y permitía que la unidad leyera y escribiera datos en él. Una pequeña abertura en un lateral de la carcasa hacía el papel de protección contra escritura. Dicha protección se controlaba pegando o no un adhesivo sobre la abertura.

El disco de 5,25" presentaba varias desventajas. No sólo era un dispositivo muy lento y con muy poca capacidad de almacenamiento (de 100 kB hasta 1,2 MB), sino que era extremadamente frágil. La carcasa de plástico hacía un buen papel de protección: el simple hecho de colocar una etiqueta sobre el disco y escribir con un bolígrafo era suficiente para dañarlo. La existencia de una ventana en la carcasa que exponía el disco hacía muy sencilla la pérdida de datos. Para evitar esto último, era necesario mantener el disco en el interior de una funda de papel.

Debido a tal conjunto de problemas, el disco de 5,25" fue relevado por el formato de 3,5", introducido por Sony en los años 90 y utilizado actualmente en los PC. El concepto es muy similar, pero soluciona todos los problemas antes planteados. En primer lugar, el tamaño es mucho más reducido, lo cual lo hace más manejable. La carcasa es mucho más rígida, por lo que el disco es más resistente a golpes y permite escribir sobre ella sin causar daños. La ventana de lectura y escritura ya no expone el disco, puesto que se halla cubierta por una protección metálica corrediza. Dicha protección se desplaza en el interior de la unidad, exponiendo el disco sólo cuando es realmente necesario. Esto hace que la duración de los datos se incremente notablemente.

Además, la abertura circular que permite girar al disco ya no es un agujero, sino que presenta una pieza, metálica. Esto proporciona todavía más protección. La protección contra escritura se controla mediante una pequeña pieza de plástico, que en ningún momento deja expuesto el interior del disco, lo que también contribuye a una mayor duración de la información. Aunque muchos fabricantes proporcionan pequeñas fundas de plástico, su uso no es estrictamente necesario, gracias a todas las características comentadas.

En resumen, el disco de 3,5" es mucho más resistente, duradero y manejable que el disco de 5,25", y esa es la razón por la cual ha persistido hasta la actualidad.

El soporte de almacenamiento

Como ya se ha introducido, el disco flexible consiste en una superficie circular maleable, recubierta de material magnético. La información se almacena mediante la introducción de pulsos magnéticos sobre el disco, y se lee siguiendo el mismo principio. En realidad, dicho principio es el mismo que se utiliza en las cintas magnéticas, salvo que en el disco flexible se almacena información en ambas caras del material. Antiguamente existían discos de una sola cara, e incluso discos de dos caras que había que girar manualmente.

La superficie del disco se divide en anillos concéntricos denominados pistas. No se deben confundir con los microsurcos de un disco de vinilo, que se encuentran totalmente conectados en forma de espiral. En el caso de un disco magnético, las pistas son concéntricas, y por tanto inconexas. A su vez, hay una división radial que divide a todas las pistas en un mismo número de porciones, denominadas sectores (ver Figura 2 para una mejor comprensión).

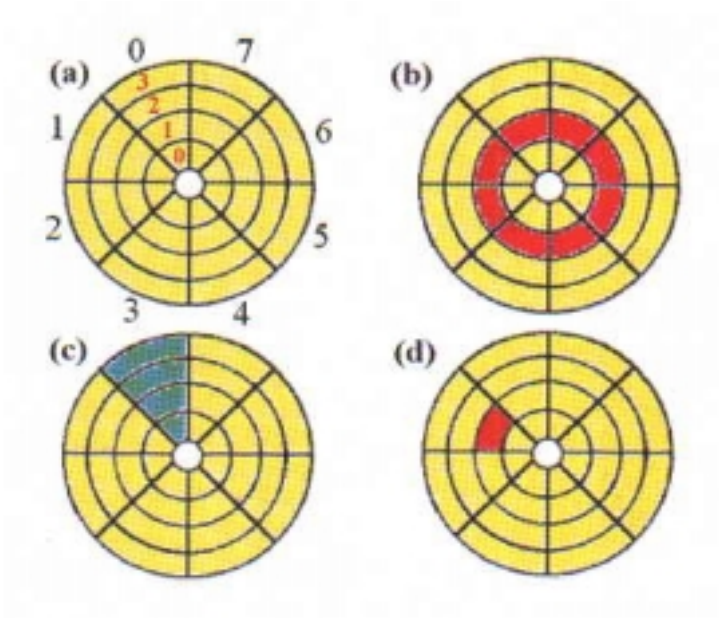


Figura 2. (a) Organización del disco en pistas (numeradas de 0 al 3) y sectores (numerados de 0 a 7); (b) Pista 1; (c) Sector 0; (d) Región de almacenamiento definida por la pista 1 y el sector 1, capaz de almacenar 512 Bytes.

En cada acceso, la unidad puede leer o escribir en el sector definido por una pista y una de las porciones. El sector es la unidad más pequeña de lectura y escritura en un disco flexible, y suele albergar 512 bytes. Dicho de otro modo, en cada acceso se leen o escriben 512 bytes (esto se hace así por razones de rendimiento). Por ejemplo, si un disco tiene normalmente 80 pistas, 18 sectores y dos caras, se obtiene que 80 pistas x 18 sectores/pista x 2 caras x 512 bytes = 1,44 MB de capacidad de almacenamiento (no olvidar que 1 MB = 1.024 bytes).

En el campo del almacenamiento en disco, un factor muy importante es la densidad superficial. Ésta mide lo concentrada que se halla la información en el disco. Por ello, si se comparan discos de igual tamaño, a mayor densidad su-

perficial, mayor capacidad de almacenamiento. La densidad superficial se calcula como el producto de otras dos densidades: la densidad de pistas y la densidad lineal.

La densidad de pistas indica la cantidad de pistas que existen por unidad de longitud, es decir lo "apretadas" que están las pistas. En esta definición, la longitud se mide en sentido radial, desde el centro del disco, y las unidades son pistas por pulgada (PPI). Otra posible medida es la densidad lineal, que informa sobre lo comprimida que se halla la información dentro de las pistas (es decir, lo "apretados" que están los bits en cada pista). La densidad lineal se expresa en bits por pulgada por pista (BPI). La multiplicación de ambas densidades da lugar a la densidad superficial, que se mide en bits por pulgada cuadrada.

En función de la densidad, existen dos especificaciones estándares: doble densidad (DD) y alta densidad (HD). Los discos DD de 5 25" tenían una densidad de 48 PPI, permitiendo almacenar tan sólo 360 kB. La variante HD en 5,25" proporcionaba 1,2 MB de almacenamiento, utilizando 96 PPI.

En el caso de los discos de 3,5", se obtenían 720 kB en DD, con 135 PPI. La variante HD obtenía 1,44 MB utilizando también 135 PPI. Entonces, ¿por qué ofrece el doble de capacidad? La respuesta es sencilla: porque duplica la densidad lineal.

En el caso de los discos de 3,5" aún existe una especificación más: densidad extra-alta (ED). En este caso, se alcanza una capacidad de 2,88 MB, también con 135 pistas por pulgada (duplicando la densidad lineal respecto a la variante HD).

Llegados a este punto, es importante indicar que con los discos HD se usan señales más suaves y dotadas de características diferentes a las de los discos DD. Se puede formatear un disco DD como HD (basta con realizar una perforación en la parte inferior derecha del disco), pero el soporte no está preparado para dicho tipo de almacenamiento y la información resultará menos duradera de lo esperado.

Unidades de disco flexible

La Figura 3 muestra el aspecto interno de una unidad de disco flexible. El aspecto externo queda reflejado en la Figura 4.

Un elemento fundamental son los cabezales de lectura y escritura, que se encargan de leer y almacenar la información, interaccionando magnéticamente con el disco. Existen dos cabezales que se mueven de forma solidaria: uno de ellos cubre el disco por la parte superior, y el otro por la parte inferior. De esta forma es posible leer o escribir ambas caras sin necesidad de extraer el disco y girarlo.

Los cabezales se montan sobre una pieza móvil, que forma parte del componente denominado "actuador". Este dispositivo se encarga de colocar los cabezales sobre la pista apropiada del disco, y se basa en un motor paso a paso (que realiza movimientos entre posiciones predefinidas). Aún queda algo por resolver: el actuador permite seleccionar la pista, pero ¿cómo se selecciona el sector? En lugar de mover los cabezales hacia el sector deseado, resulta mu-

cho más simple hacer girar el disco, mediante un motor. Todavía surge una pregunta: ¿cómo se determina cuál es el sector que está debajo de los cabezales en cada momento? Para ello se utiliza un punto de referencia, que consiste en una perforación realizada sobre el propio disco. Utilizando un sensor óptico, es sencillo averiguar el momento en que la perforación pasa por debajo del mismo (ya que deja pasar la radiación luminosa), y por tanto tomar un punto de referencia estable para localizar los sectores. La Figura 5 muestra el aspecto de los cabezales de lectura y escritura.

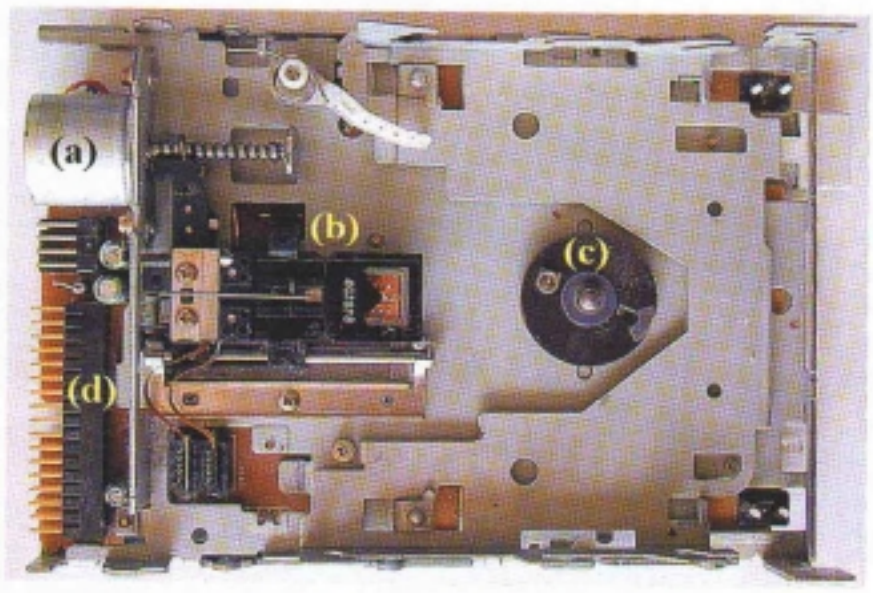


Figura 3. Estructura interna de una unidad de disco flexible. (a) Motor paso a paso, encargado de desplazar los cabezales de lectura y escritura. (b) Cabe-zales de lectura y escritura. (c) Motor giratorio. (d) Interfaz de conexión con el resto del sistema y alimentación

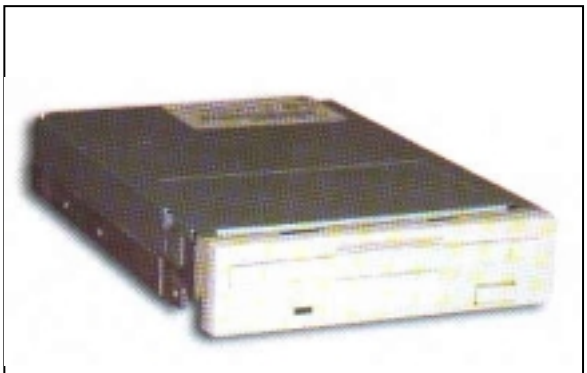


Figura 4. Aspecto externo de la unidad de disco flexible

Un hecho importante (y que diferencia a los discos flexibles de los duros) es que los cabezales establecen contacto real con el disco. Es por ello que los discos flexibles giran a una velocidad modesta, usualmente de 360 revoluciones por minuto (RPM). Si giraran más rápido, el disco resultaría dañado debido al contacto físico.

Junto con los cabezales se encuentra una bobina que, antes de escribir información en un sector, elimina su contenido. El borrado toma una anchura superior a la del sector, de forma que se asegura que los sectores presentes en las pistas adyacentes no van a interferir.

Otro componente importante es el sensor de cambio de disco. Dicho sensor envía una señal al controlador de la unidad cuando el usuario extrae un disco y lo

reemplaza por uno nuevo. Esto mejora mucho el rendimiento, ya que el sistema no debe examinar continuamente si el contenido del disco ha cambiado. El sistema almacena el contenido del disco en una caché (porción de la memoria RAM), y accede a ella en lugar de al disco. Tan solo habrá que refrescar la caché cuando el sensor indique que el disco ha cambiado.

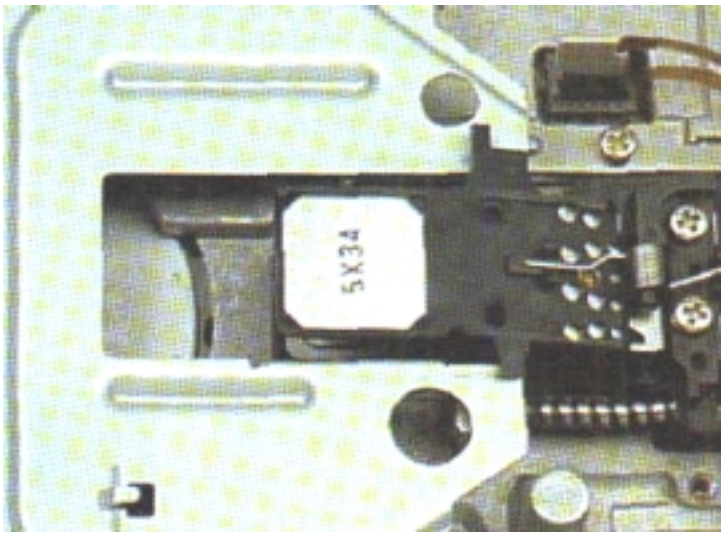


Figura 5. Cabezales de lectura y escritura. Nótese que el disco quedará cubierto por ambas partes

No podía faltar un componente imprescindible: el hardware controlador que -como se introdujo en el capítulo anterior para los dispositivos IDE- se encuentra integrado en la propia unidad. Después de todo lo comentado hasta ahora, resulta más sencillo comprender las funciones de los circuitos controladores. En el caso del disco flexible, el controlador es el encargado de actuar sobre los elementos antes vistos (cabezales, actuador, etc.) de la forma apropiada y en el orden correcto. Además, debe proporcionar una interfaz de conexionado con el resto del sistema para recibir órdenes, devolver información, y recibir las señales de alimentación.

Algunos problemas de las unidades de disco flexible

Como ya hemos comentado, los cabezales se posicionan sobre las pistas mediante un motor que tiene varias posiciones previamente establecidas. Esto conduce a ciertos problemas conocidos. Si los discos se someten a elevadas temperaturas, la dilatación hará que las pistas se desplacen de la posición estándar. Al leer o escribir, el posicionador intentará acudir a las posiciones prefijadas, pero en realidad estará trabajando sobre pistas equivocadas, conduciendo a un evidente error.

Otro problema típico es la pérdida de la alineación del posicionador. Si esto ocurre, el posicionador perderá la costumbre de buscar las pistas en las posiciones estándar, y, por tanto, acudirá a posiciones equivocadas. Si un disco se formatea y graba sobre una unidad con problemas de alineación, y se intenta leer en una unidad sana, se producirán errores. Cuando se intente acceder a una pista, se estará accediendo a una pista equivocada. La única solución para poder usar ese disco en la unidad sana es formatearlo de nuevo con esta última.

El problema de alineación lo puede resolver el fabricante. Esta práctica era común antiguamente, cuando las unidades de disco flexible eran caras. Actual-

mente, el precio es tan económico que resulta más rentable adquirir una nueva unidad.

El sensor de cambio de disco también puede generar importantes problemas. En efecto, suponga que el sensor no funciona correctamente. En ese caso, si cambiamos de disco, el sistema pensará que el disco anterior aún permanece en la unidad. Al intentar leer un archivo, el sistema indicará que el archivo no se ha encontrado. La escritura es aún más peligrosa, puesto que corromperá el contenido del nuevo disco, ya que el sistema cree que escribe en el anterior.

Formateo de discos

El proceso de formateo prepara un disco para que la información pueda ser almacenada y leída. La costumbre de emplear tan sólo un programa (el conocido FORMAT.COM de MS-DOS) ha conducido frecuentemente a pensar en un único procedimiento. Pero, en realidad, existen dos procesos diferentes: el formato a bajo nivel y el formato a alto nivel. Como se verá en la próximo capítulo, el formateo de discos duros implica un tercer procedimiento (creación de particiones).

El formato a bajo nivel crea sobre el disco las estructuras que almacenarán la información (pistas y sectores). Esto implica crear las pistas y definir el inicio de los sectores dentro de cada pista. Este tipo de formato se conoce también como "formato real" ya que se registra realmente la organización del disco.

El formato a alto nivel crea las estructuras lógicas que empleará el sistema operativo, tales como la tabla de asignación de archivos (FAT) y el directorio raíz. De cara al usuario, la organización en pistas y sectores no resulta útil para trabajar con el disco. En cambio, la organización que el usuario desea emplear se basa en archivos. Es ahí donde el sistema operativo presta un servicio al usuario, haciendo ver archivos (estructuras de alto nivel) y ocultando la complejidad de la organización real (pistas y sectores, estructuras de bajo nivel). La FAT se emplea para realizar la traducción entre bajo y alto nivel, definiendo el sistema de archivos. No es más que una tabla que almacena la posición de los sectores que componen cada archivo, en su correcto orden (los archivos no tienen por qué ocupar sectores contiguos). La FAT se aloja en los primeros 63 sectores del disco.

Si se reformatea un disco a alto nivel, dicha tabla se vacía, de forma que el disco parece estar vacío de cara al usuario. Pero, en realidad, la información grabada en el disco sigue intacta. Por ello -si se emplean las herramientas adecuadas- es posible deshacer el formato y devolver el disco a su estado anterior.

Mientras que el formato a bajo nivel se mantiene, el formato a alto nivel varía de unos sistemas operativos a otros. Cada sistema operativo tiene definido su propio sistema de archivos, empleando estructuras con diferentes características. Por tanto, cada sistema realizará el formato a alto nivel de una forma distinta (empleando programas de formateo diferentes). Como consecuencia, si se adquieren discos preformateados, se debe comprobar para qué sistema operativo son válidos, o será necesario formatearlos de nuevo.

CAPÍTULO 9

UNIDADES Y SOPORTES DE ALMACENAMIENTO: DISCOS DUROS, ZIP Y JAZ

Este capítulo se centra en el funcionamiento del disco duro presentando su historia, estructura, funcionamiento y variantes existentes. Además, hablaremos de dos medios de almacenamiento magnético realmente actuales: los discos ZIP y JAZ.

Detalles generales del disco duro

En esta ocasión vamos a abordar un medio de almacenamiento interno y fijo. Ya no hay una unidad fija al PC y un soporte de almacenamiento extraíble, sino que ambas partes se encuentran en el interior del PC. De cara al usuario, el PC dispone de una unidad de gran capacidad y velocidad, sin ocupar espacio exterior, sin necesidad de emplear cables, y que no requiere de un soporte de almacenamiento extraíble. El disco duro va con el PC a todas partes, almacenando los datos vitales para los programas y el sistema operativo.

La tecnología en el campo de los discos duros ha demostrado un continuo y asombroso avance desde sus inicios en los años 50. El avance tecnológico ha apuntado siempre hacia la mejora de dos parámetros: mayor capacidad y velocidad. También se ha perseguido la reducción de tamaño, aunque en un nivel de importancia inferior a los parámetros anteriores.

Para conseguir mayor capacidad, la lucha consiste en obtener mayores densidades superficiales de información. El aumento de la velocidad se consigue aumentando la velocidad de giro del disco, que viene condicionada principalmente por las características de los cabezales de lectura y escritura. Tal y como se ha introducido, estos parámetros no han dejado de mejorar y continúan haciéndolo a una velocidad asombrosa. Otro parámetro en constante evolución es el precio por MB, que decrece también de forma asombrosa con el tiempo: los discos duros son cada día más rentables.

El avance de los discos duros tiene un importante impacto en el rendimiento del PC. En primer lugar, los programas (empezando por el sistema operativo) son cada día más voluminosos y acceden a mayores cantidades de datos. Esto exige capacidad de almacenamiento (para almacenar los programas y los datos), además de velocidad (para agilizar el acceso a dichos datos). Por otro lado, el arranque del PC será más rápido cuanto más veloz sea el disco duro.

Otro punto importante radica en la capacidad multitarea de los sistemas operativos actuales. Cuando se ejecutan muchos procesos simultáneos, es probable que no haya suficiente memoria para albergarlos a todos. Lo mismo ocurre si no son muchos los procesos, pero consumen grandes cantidades de memoria. En esos casos, la memoria RAM no proporciona suficiente espacio de almace-

namiento, y se utiliza el disco duro como memoria virtual. Si el disco duro no es suficientemente rápido y no dispone de mucho espacio libre, el usuario apreciará que sus programas se ejecutan lentamente y que el sistema operativo apenas responde.

Un poco de historia

Tal y como hemos introducido anteriormente, el disco duro inició su carrera en los años 50, y no ha parado de avanzar hasta la actualidad (de hecho, sigue avanzando). La firma IBM ha jugado un papel fundamental en dicha evolución desde el primer momento.

Figura 1. Los inicios: el disco duro IBM 305 RAMAC

En efecto, el primer disco duro fue desarrollado por IBM en 1956. Recibió el nombre de RAMAC (Figura 1), y constaba internamente de 50 discos de 24" cada uno. La capacidad total ofrecida era de 5 MB.

Sin embargo, el padre del disco duro moderno nació en 1973, también de la mano de IBM. Su nombre era 3340, y constaba de dos módulos de 30 MB, uno fijo y el otro extraíble. Estableciendo símiles entre algunas de sus características y las de un conocido rifle, fue bautizado con el apodo "Winchester". Aunque mucho más avanzados, los discos duros actuales se basan totalmente en los conceptos introducidos en aquel disco duro. Uno de los conceptos principales radica en que las cabezas de lectura / escritura son flotantes (es decir, no existe contacto físico con la superficie del disco). De hecho -además de aumentar la densidad superficial- una parte fundamental del avance consiste en optimizar la distancia entre las cabezas y el disco sin llegar al contacto.

La entrada del disco duro en el mundo del PC se produjo con el lanzamiento de la variante XT. Se incorporaba un disco duro ST-412 de 10 MB, fabricado por Seagate.

Arquitectura del disco duro

La estructura interna del disco duro queda ilustrada en la Figura 2. Básicamente, el disco duro está integrado por un conjunto de discos de igual diámetro, comúnmente denominados "platos". Cada plato se compone de un sustrato de elevada rigidez, que se recubre con un material magnético. El nombre de disco duro proviene, precisamente, del alto grado de rigidez de los platos (en oposición a lo que ocurría con los discos flexibles)
Los platos se hallan montados sobre un eje, y se mantiene una cierta distancia entre ellos, gracias a unos anillos separadores. El número usual de platos osci-

la entre 1 y 4 en discos duros normales. Los discos duros de alta capacidad pueden llegar a incorporar más de 10 platos.

El eje se halla gobernado por un motor giratorio. Cuando el motor gira, el eje gira, y, por tanto, todos los platos giran a la misma velocidad.

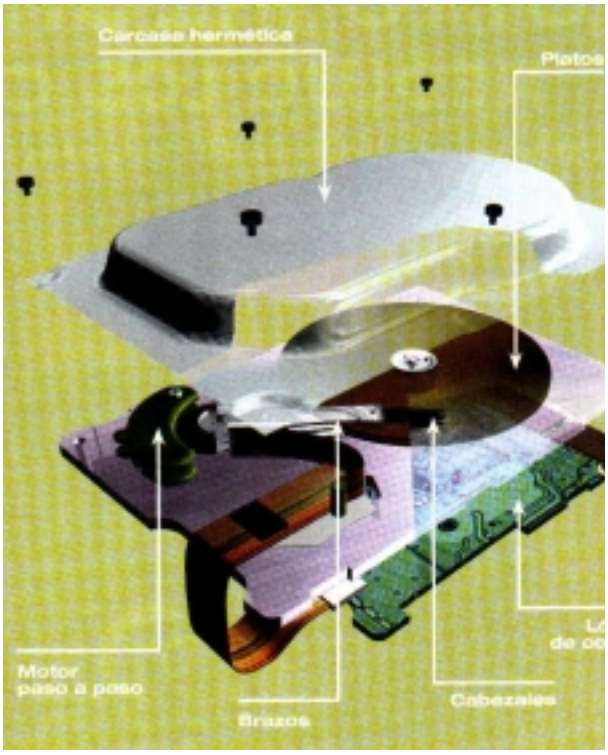


Figura 2. Estructura interna del disco duro

Los elementos encargados de leer y escribir la información se denominan -al igual que ocurría en los discos flexibles- cabezales de lectura y escritura. Estos se encargan de convertir bits en pulsos magnéticos (al escribir) o bien pulsos magnéticos en bits (al leer). Hay dos cabezales dedicados a cada plato. Uno de ellos se sitúa en la parte superior, mientras que el otro se sitúa en la cara inferior. De esta forma es posible acceder de manera rápida a ambas caras de cada plato. Ya que el número usual de platos oscila de 1 a 4, el número habitual de cabezales oscilará entre 2 y 8.

Como ya hemos indicado antes, la diferencia principal respecto a los discos flexibles radica en que los cabezales no tocan la superficie de los platos. Esto permite que el disco gire a mayor velocidad, generando menos calor y produciendo menos nivel de ruido. Mayor velocidad de giro significa menor tiempo de acceso a la información, y por tanto mayor velocidad de trabajo.

Conviene lograr que los cabezales se encuentren a una distancia óptima de los platos. Dicha distancia está relacionada con la potencia de las señales emitidas por los cabezales y por el disco. Si se trabaja con señales suaves, los cabezales deberían estar cerca de los platos. En otro caso, las señales no se recibirían correctamente por los cabezales al leer, ni quedarían bien registradas en los platos al escribir. En el lado opuesto, si se trabaja con señales fuertes, los cabezales deberían estar más alejados de los platos. La potencia de las señales está altamente condicionada por la densidad de la información. A mayor densidad, los bits se hallan más cercanos entre sí en los platos, y, por tanto, se requieren señales más suaves para evitar interferencias. Por ello, se deduce que a mayor densidad superficial, es necesaria una menor distancia entre cabezales y platos.

Los cabezales de lectura y escritura se montan sobre unos elementos denominados "deslizadores". Estos presionan a los cabezales sobre los platos cuando el disco está parado. Cuando el disco gira, el flujo de aire desprendido hace que los deslizadores se desplacen, colocando a los cabezales a la distancia

apropiada. La Figura 3 muestra con más detalle el aspecto de los cabezales y los deslizadores.

Los deslizadores se montan sobre unos elementos rígidos denominados brazos. Los brazos se unen a un eje, controlado por un motor paso a paso. Por tanto, los brazos se mueven solidarios. Esto significa que todos los cabezales siempre se moverán en conjunto, encontrándose siempre uno encima del otro.

Los elementos internos del disco duro se gobiernan mediante un circuito controlador, que, además, se encarga de comunicar al disco duro con el resto del PC.

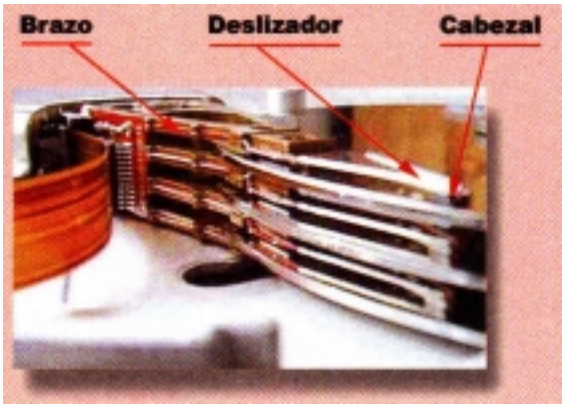


Figura 3. Brazos, deslizadores y cabezales

Es importante destacar la existencia de una memoria caché que actúa como almacenamiento intermedio pra agilizar las transferencias entre disco duro y PC (y viceversa).

Para lograr un buen funcionamiento, el disco duro exige un alto nivel de precisión en su interior. Ante todo, se debe evitar a toda costa la entrada de partículas de polvo, que dañarían los cabezales con facilidad. Por ello, el interior del disco duro se aísla fuertemente del exterior, y los componentes se ensamblan en condiciones especiales (que aseguran un ambiente totalmente libre de polvo).

Organización de la información

Las estructuras de bajo nivel empleadas en los discos duros para almacenar información son una ampliación de las utilizadas en los discos flexibles.

Al igual que ocurría con los discos flexibles, la superficie de cada plato queda dividida en pistas y sectores. La división es idéntica para todos los platos.

Los sectores siguen almacenando la misma cantidad de información: 512 bytes. Hay que anotar que dicha cantidad es realmente mayor. Normalmente se almacenan bytes adicionales, que se emplean para apoyar en el control de la unidad, y para la detección y corrección de errores. La disposición y utilización de estos bytes adicionales no sigue ningún estándar, y varía de un disco duro a otro. Cuantos más bytes adicionales se empleen, menor espacio efectivo quedará para el almacenamiento.

En el caso del disco duro, todavía hay estructuras de mayor nivel. En efecto, los sectores contiguos se agrupan formando "clusters" (agrupaciones). De hecho, el disco duro toma el cluster como la unidad más pequeña de almacenamiento. En cada acceso, se lee o escribe un cluster. Al trabajar con bloques de información más grandes, el rendimiento queda afectado de forma positiva. Los clusters no tienen un tamaño estándar. Dicho tamaño depende de varios factores, y principalmente lo decide el sistema operativo.

Otra estructura de alto nivel son los denominados cilindros (ver Figura 4). Como ya se ha introducido, los cabezales se mueven en conjunto, al estar guiados por brazos solidarios. Cuando un cabezal está sobre una pista, el resto de los cabezales está sobre la misma pista, a través de los diferentes platos (y caras) que componen el disco duro. Si imaginamos una disposición de anillos (pistas) situados uno sobre otro, obtenemos el esqueleto de un cilindro, y de ahí el nombre. Por tanto, decir que el disco está trabajando sobre el cilindro 3 significa que todos los cabezales están sobre la pista 3 de cada plato. Si un disco tiene 4 platos, tendrá 8 cabezales, y, por tanto, 8 pistas en cada cilindro.

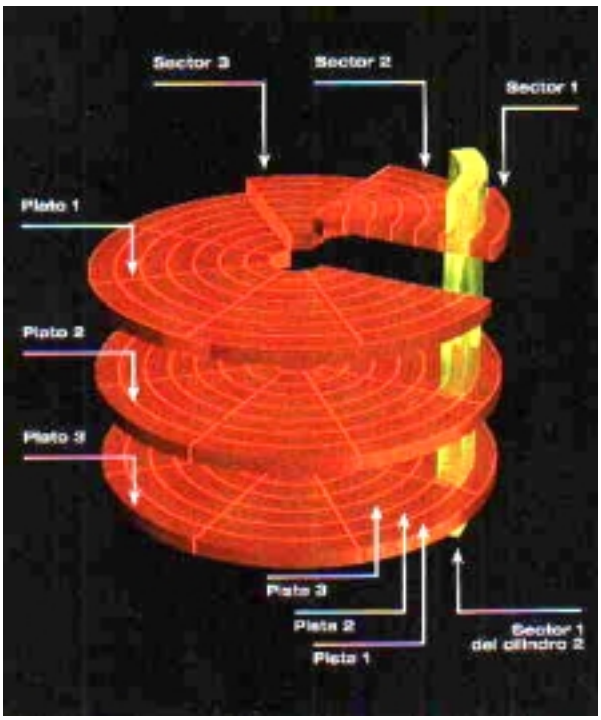


Figura 4. Cilindros, pistas y sectores

Finalmente, la estructura de mayor nivel son las particiones, que no son más que grupos de cilindros contiguos. El disco se divide en varias particiones, que el sistema operativo hace ver como unidades lógicas diferentes. Aunque se trata del mismo disco, el usuario aprecia varias letras de unidad, y cree estar trabajando con varios discos duros de menor tamaño. Una de las ventajas de las particiones consiste en que los cabezales se deberán mover dentro de un grupo conexo de cilindros de menor tamaño, y, por tanto, deberán realizar menor recorrido para encontrar el cilindro deseado en cada acceso. Esto se traduce en una mayor velocidad de acceso a la información.

Llegados a este punto, es importante comentar cómo se direcciona la información en el disco duro. En un disco flexible, se empleaban dos coordenadas: pista y sector. En el caso del disco duro, pasamos al mundo tridimensional: la información se direcciona mediante la terna (cilindro, cabezal, sector). Una vez seleccionado un cilindro, hay que seleccionar cuál es la pista deseada dentro del mismo (esto es, seleccionar un cabezal). Dentro de dicha pista, se selecciona el sector deseado.

En el campo de la organización de la información, la principal diferencia respecto al disco flexible es una mayor densidad superficial. En efecto, la densidad de pistas y la densidad lineal son extremadamente mayores.

Para optimizar la densidad superficial, algunos discos duros emplean una técnica denominada "registro por zonas". Esta técnica se basa en un hecho sencillo: cuanto más al exterior del plato se encuentra una pista, mayor longitud presenta. Como, normalmente, todas las pistas almacenan el mismo número de bits, está claro que la información estará más "comprimida" en las pistas interiores que en las exteriores. El registro por zonas intenta aprovechar mejor el espacio, tratando de igualar la densidad de bits entre las diferentes pistas. Por

tanto, se colocarán más sectores en una pista cuanto más hacia al exterior del plato se encuentre dicha pista.

Funcionamiento del disco duro

En cada acceso a la información, el disco duro realiza un amplio conjunto de tareas. A continuación indicamos los pasos fundamentales implicados en cada acceso a disco.

En primer lugar, se realizan varias etapas de traducción, que hacen que una dirección (un número que apunta a una posición del disco) se traduzca en una localización geométrica, de tipo (cilindro, cabezal, sector). A continuación, se hace girar el disco, si es que éste no estaba ya en marcha. Cuando el disco alcanza una velocidad estable, se mueven los cabezales hacia el cilindro deseado, utilizando el motor paso a paso que controla los brazos. Una vez en el cilindro apropiado, se activa el cabezal correspondiente al plato deseado. Entonces, se espera el tiempo necesario para que el giro del disco haga pasar al sector deseado bajo el cabezal. Cuando esto ocurre, se lee o se escribe la información en dicho sector.

Como ya hemos adelantado, este proceso sólo ha incluido los pasos fundamentales. En realidad el proceso implica varios pasos adicionales. Por ejemplo, no hay que olvidar la existencia de una caché que acelera las transferencias.

El rendimiento del disco duro está determinado por la eficiencia con que se realizan estos pasos. Principalmente, el tiempo de acceso a la información es la suma del tiempo empleado en llevar los cabezales hacia el cilindro adecuado y el tiempo de giro del disco hasta encontrar el sector buscado. El tiempo de acceso total suele oscilar entre 10 y 20 milisegundos.

Formateo del disco duro

En el caso del disco duro, los fundamentos del formateo son análogos a los introducidos para los discos flexibles. El formateo de un disco duro presenta tres pasos fundamentales. En primer lugar, se da formato a bajo nivel, tal y como explicamos en la anterior entrega. A continuación, se procede a la creación de particiones (proceso introducido anteriormente en este artículo). Finalmente, se procede al formateo a alto nivel (también explicado en la anterior entrega). No hay que olvidar que el conocido comando FORMAT de MS-DOS se comporta de forma diferente según se trabaje con discos duros o discos flexibles.

Variantes del disco duro

Hasta ahora se ha hablado del disco duro como elemento interno y fijo, ya que ésta es la variante estándar en cualquier PC. Sin embargo, conviene comentar la existencia de otras variantes del disco duro: los discos duros externos y los extraíbles. Estas variantes funcionan de forma idéntica, presentando tan sólo diferencias en cuanto a su montaje en el PC.

Los discos duros externos constituyen un dispositivo portátil que trabaja en el exterior del PC. Vienen cubiertos por una carcasa de plástico, y presentan una

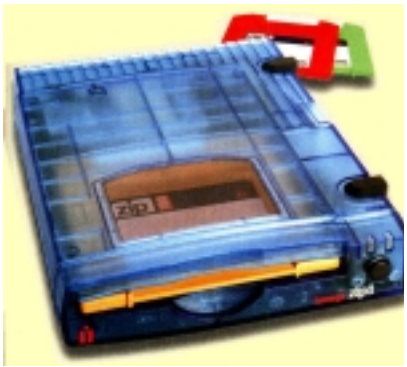
fuente de alimentación. Estos discos son más caros que los internos, pero son más fáciles de instalar, presentan mejor refrigeración, y son más fáciles de expandir e interconectar con otros dispositivos. El hecho de no tener que adaptarse al tamaño de un PC dota al fabricante de mayor libertad de diseño, lo que siempre reporta ventajas. Estos dispositivos se suelen conectar al PC mediante el uso de la interfaz SCSI.

Los discos duros extraíbles presentan un buen compromiso entre los discos duros internos y los externos. En este caso, se instala un módulo fijo en una de las bandejas estándar del PC para la instalación de unidades de disco. Sobre dicho módulo se puede introducir un módulo extraíble, que es el disco duro en sí. En el fondo, se trata de utilizar el disco duro como si se tratara de un soporte de almacenamiento extraíble, al estilo de un disco flexible.

Discos JAZ

Los discos JAZ se basan en otra tecnología desarrollada por Iomega. Básicamente, se trata de tomar un disco duro y colocarlo dentro de una carcasa. Dentro de la carcasa se encuentran varios platos, pero no se hallan los cabezales, el motor paso a paso, los brazos, etc. Todos esos elementos se encuentran en la unidad instalada en el PC. Por ello, no hay que confundirlos con los discos duros extraíbles, que contienen todos los elementos de un disco duro dentro de la carcasa. Evidentemente, un disco JAZ presenta un excelente rendimiento y una gran capacidad, precisamente debido a su condición de disco duro. Sin embargo, al perder la estanqueidad, no es posible conseguir las mismas características que en un disco duro hermético. Actualmente, existen discos JAZ en versiones de 1 GB y 2 GB.

Unidad de disco JAZ



Unitat ZIP

Discos ZIP

Los discos ZIP (desarrollados por la firma Iomega) se presentan como un medio de almacenamiento magnético similar a los discos flexibles, pero con una capacidad notablemente superior. La tecnología empleada recibe el nombre de *Bernoulli*, también introducida por Iomega. La principal diferencia respecto a un disco flexible se centra en el recubrimiento magnético del disco. En un disco ZIP, dicho

recubrimiento es de alta calidad, lo que se traduce en una gran densidad superficial de información. Los cabezales se fabrican con un tamaño bastante más reducido, de forma que puedan afrontar tal densidad de información. El mecanismo de posicionamiento de los cabezales es similar al de los discos duros. Además, se utiliza la técnica de registro por zonas, introducida en este mismo artículo. Gracias a todos estos detalles, es posible encontrar discos ZIP (y sus correspondientes unidades) de 100 MB y 250 MB. Las unidades de 250 MB son capaces e leer también los discos ZIP de 100MB. En cuanto al rendimiento, es muy superior al de un disco flexible, ya que todos los detalles expuestos le permiten alcanzar una velocidad de rotación notablemente más alta: 3.000 rpm.

CAPÍTULO 10

EL CD Y DVD

Básicamente, los discos ópticos son un soporte digital de almacenamiento de datos basado en el comportamiento de la radiación luminosa. La primera generación de discos ópticos fue introducida por la firma Philips, en colaboración con Sony, a finales de los años 80. La tecnología asociada a dichos discos fue denominada CD (Compact Disc), y se ha convertido en el soporte más utilizado actualmente para el almacenamiento y distribución de audio y software (entre otros tipos de información). En este artículo vamos a recorrer la tecnología CD, prestando atención a las variantes más relacionadas con el mundo del PC: CD-ROM, CDR y CD-RW. También presentaremos la tecnología DVD.

La tecnología CD

Un disco basado en la tecnología CD es capaz de almacenar 74 minutos de audio (650 MB de datos y programas), e incluso existen variantes capaces de almacenar hasta 99 minutos (870 MB de datos y programas).

En principio, la grabación de un CD se realiza en fábrica. Antes de comenzar, se utiliza un láser muy potente para realizar perforaciones en un disco maestro. Dichas perforaciones son las que almacenan la información digital. A partir del disco maestro se genera un molde, que se utilizará para grabar copias.

El material básico que compone un CD es una pieza de policarbonato circular de 1,2 mm de espesor Empleando el molde, se "copian" las perforaciones (antes realizadas sobre el disco maestro) sobre la pieza de policarbonato. En otras palabras, el grabado se hace "a presión", y no mediante un láser como ocurría en el disco maestro. Las zonas perforadas se denominan "huecos" (más conocidos como *pits*) y las zonas sin perforar se denominan *lands*.

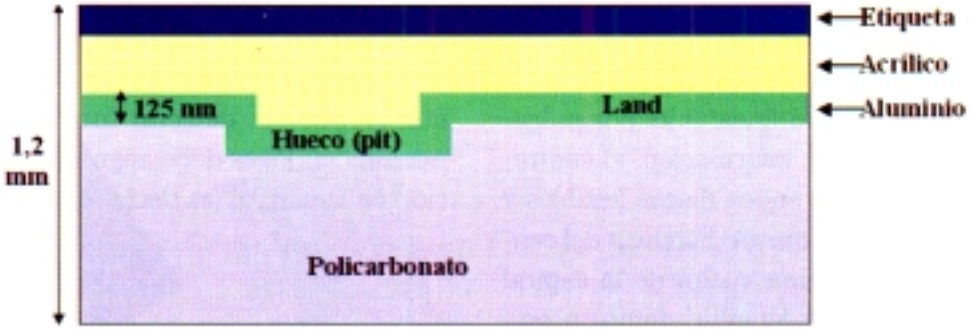


Figura 1. Corte transversal de un CD

Tras la perforación, el bloque de policarbonato se recubre de una delgada capa de aluminio, de tan sólo 125 mm de espesor (0,125 micras). Sobre ésta se deposita otra capa de acrílico transparente, que sirve como protección. Finalmente, se imprime la etiqueta del CD sobre la capa de acrílico.

La Figura 1 muestra un corte transversal del CD, donde se puede apreciar el efecto de las perforaciones (la existencia de *lands* y *pits*) y las diferentes capas descritas.

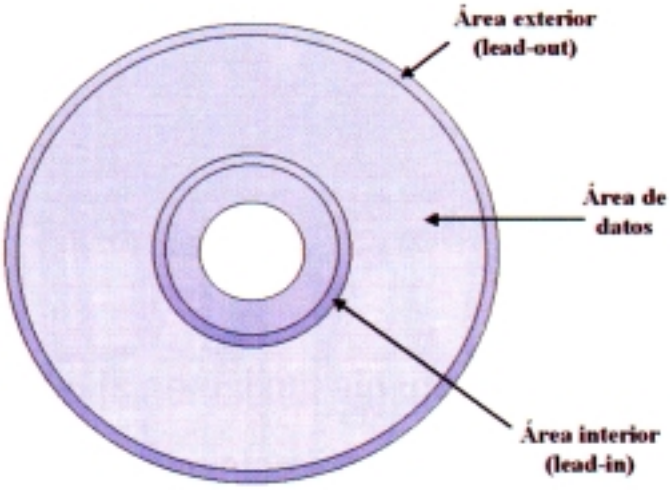


Figura 2. Estructura de un CD

Una vez producido, un disco CD tiene la estructura mostrada en la Figura 2. Se pueden distinguir tres áreas principales. El área interior (*lead-in*) contiene dos canales de información. Uno de ellos contiene silencio digital (estados lógicos 0) y el otro la tabla de contenido del CD. A esta zona le precede información para la alineación del láser de lectura con el comienzo de la zona de datos.

El área de datos contiene la información grabada en el CD (música, programas, vídeo, etc.). Finalmente, el área exterior (lead-out) contiene silencio digital, indicando el final del CD.

Lectura de un CD

La lectura de un CD consiste en la conversión de los *lands* y *pits* a información digital (ceros y unos). El elemento fundamental para la lectura de un CD es un láser de baja potencia, que emite radiación infrarroja y que se enfoca hacia la parte inferior del CD. La luz atraviesa la capa de policarbonato e incide sobre la capa de aluminio. Si el haz incide sobre un hueco (*pit*), el porcentaje de luz reflejada es muy pequeño. Por el contrario, si el haz incide sobre una zona plana (*land*), un gran porcentaje de luz es reflejada. La radiación luminosa reflejada se dirige hacia un fotodetector que, en función de la intensidad de la luz recibida, puede detectar fácilmente si se ha enfoca-do un *land* o un *pit*.

La transformación de *lands* y *pits* a valores digitales no sigue una correspondencia directa. En otras palabras, un *land* no significa un valor digital "0", y un pit no significa un valor digital "1". En realidad, un *land* indica mantener el estado digital anterior, y un *pit* indica invertir el estado anterior. Con esto se consigue minimizar la cantidad de perforaciones necesarias sobre el CD, lo que permite grabar un CD más rápidamente. La Figura 3 muestra un ejemplo de paso de *lands/pits* a valores digitales.

Un CD no contiene pistas concéntricas, como ocurría en los discos magnéticos. En cambio, el CD presenta una sola pista, que se dispone en forma de espiral, cubriendo toda el área de datos. La espiral comienza en la parte interior del disco, justo después del área interior. Esto se hace así para permitir recortar el radio del CD y poder obtener versiones más pequeñas (como son los conocidos *CD-Single*). La anchura de la espiral es sumamente fina: tan sólo 0,5 micras. La separación entre vueltas es, de nuevo, muy reducida: 1,6 micras.

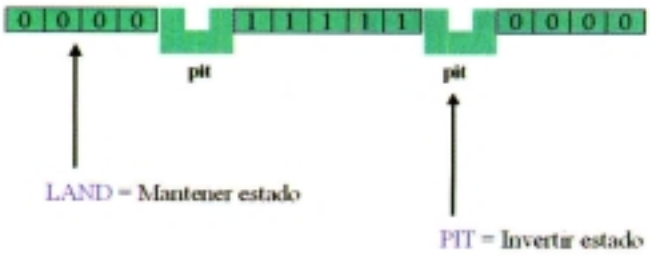


Figura 3. Ejemplo de interpretación de zonas planas y huecos, obteniendo estados lógicos como resultado

Esto explica que un solo CD pueda almacenar cantidades de información tan elevadas. Para ilustrar lo que esto significa, basta decir que si se "desenrollara" la espiral hasta convertirla en una línea recta, tendría una longitud de unos 5 kilómetros.

Cada vuelta de la espiral presenta la misma densidad lineal (bits por pulgada por pista) de información, al contrario que ocurría en los discos flexibles y duros. Como a mayor distancia del centro del disco, una vuelta de la espiral presenta mayor longitud, también presentará mayor cantidad de información. Esto proporciona un mejor aprovechamiento del espacio del disco, consiguiendo mayor capacidad de almacenamiento.

En el caso de los discos magnéticos se aplicaba una velocidad angular constante (conocida como CAV) al rotar los discos. En cambio, los CD se hacen girar a velocidad lineal constante (CLV). Esto se hace para que la información se lea a "ritmo" constante. Si el disco fuera CAV, las vueltas de la espiral más alejadas del centro girarían con mayor velocidad lineal. Visto de otro modo, si el disco gira un cierto ángulo, el número de bits comprendidos en dicho ángulo es mayor conforme la espiral se aleja del centro. Esto significaría que se producirían más bits por segundo conforme la espiral se aleja del centro. Para compensar este efecto, la velocidad angular de giro del disco se reduce proporcionalmente conforme la espiral se desplaza hacia el exterior. En el ámbito práctico, la velocidad angular varía desde 500 RPM (en la parte interior del disco) hasta 200 RPM (en la parte exterior).

El dispositivo lector de un CD tiene un diseño sencillo, en comparación con las unidades de disco duro. Pero -como hemos podido comprobar- las dimensiones con las que se almacena la información en un CD son extremadamente reducidas. Esto se traduce en que el dispositivo lector de CD deberá caracterizarse por mecanismos de una precisión extrema.

En primer lugar, el lector debe presentar un motor giratorio cuya velocidad angular se pueda variar con elevada precisión (el incremento de velocidad a aplicar es extremadamente pequeño entre vueltas de la espiral).

El láser se debe focalizar con suma precisión sobre la espiral, de forma que se puedan detectar lands y pits. Además, el láser debe seguir la evolución de la espiral (es decir, aplicar desviaciones de 1,6 micras con cada vuelta de la espiral). Por ello, el láser se monta sobre un dispositivo deslizante denominado "sistema de seguimiento". Éste desplaza el láser desde el centro del disco hacia el exterior, siguiendo la espiral con suma precisión (no hay que olvidar que este sistema debe desplazarse con una precisión del orden de la micra).

Tras la detección de lands y pits mediante el fotodetector, se encuentra el hardware que traduce dicha información a datos digitales. Dicho hardware varía según el tipo de información almacenada (por ejemplo, en el caso de la información audible, es necesaria una conversión digital /analógica).

Estructuras de datos

Tal y como ocurría con los discos magnéticos, no resulta útil acceder a la información en forma de bits. En cambio, es más eficaz agrupar la información en estructuras de alto nivel. Dichas estructuras (totalmente análogas a los sectores) tienen diferentes formatos dependiendo del tipo de información almacenada. Por ejemplo, en el caso del almacenamiento de audio, multitud de bits se agrupan bajo estructuras que incluyen datos como el instante de reproducción (horas, minutos y segundos), e incluso información para detectar y corregir errores. Dichas estructuras se agrupan en divisiones lógicas de mayor nivel, denominadas pistas. En un CD de audio, cada pista suele corresponder a un tema musical. Pueden existir hasta 99 pistas en un CD.

No es nada improbable que un CD salga de fábrica con errores en la información almacenada, lo que significa que algunos bits pueden tener su estado lógico invertido. Por ello, las estructuras de almacenamiento de información suelen incluir códigos para la detección de errores (denominados EDC) e incluso para la corrección de los mismos (denominados ECC) si la pérdida de información es crítica. Por ejemplo, en el almacenamiento de software no se puede permitir ningún error, y por tanto es necesaria la existencia de códigos EDC y ECC. En el caso de audio o vídeo, un error no es tan grave y basta con detectarlo. El reproductor puede emplear diversos algoritmos (como la interpolación) para predecir el valor real de los bits perdidos y, de esta forma, el usuario no percibe la existencia del error.

Soportes basados en la tecnología CD

La tecnología CD ha dado lugar a diferentes variantes de soporte de almacenamiento. Todas siguen los principios de dicha tecnología, y se caracterizan principalmente por la naturaleza de la información almacenada y el formato empleado al almacenar los datos (que afecta directamente a la forma de recu-

DENOMINACIÓN DEL DOCUMENTO ESTÁNDAR	VARIANTES DESCRITAS	COMENTARIOS
Libro rojo	CD-audio, CD-Graphics, CD-Text	
Libro amarillo	CD-ROM y la extensión CD-ROM XA	CD-ROM XA ha dado lugar a variantes como Photo-CD y Video-CD.
Libro verde	CD interactivo (CD-i)	Almacenamiento de información multimedia y programas. El reproductor implementa un sistema operativo. El libro verde es la especificación más amplia, completa y fácil de entender.
Libro naranja	CD grabables con capacidad multi-sesión: CD-MO (soporte magneto-óptico), CD-R y CD-RW.	
Libro blanco	Video CD	Del Video-CD derivan variantes como Karaoke-CD, VCD y Super VCD.
Libro azul	CD Extra	Permite almacenar audio y datos, pero no es grabable.
Libro morado	DDCD	Discos CD-R y CD-RW de doble densidad con capacidad para 1,3 GB.

Tabla1. Variantes de la tecnología CD y documentos estándares.

Tabla 1. Variantes de la tecnología CD y documentos estándares.

perarlos). Las especificaciones correspondientes a cada variante se encuentran incluidas en documentos estándares, que reciben una denominación "por colores". La Tabla 1 explica el contenido de cada uno de los documentos existentes.

En los próximos apartados nos centraremos en las variantes más relacionadas con el almacenamiento de datos en el mundo del PC: CD-ROM, CD-R y CD-RW.

El CD-ROM

El CD-ROM es la variante de la tecnología CD empleada para almacenar datos y programas informáticos. Un CD-ROM puede almacenar hasta 870 MB de información, y todas sus características están especificadas en el "libro amarillo".

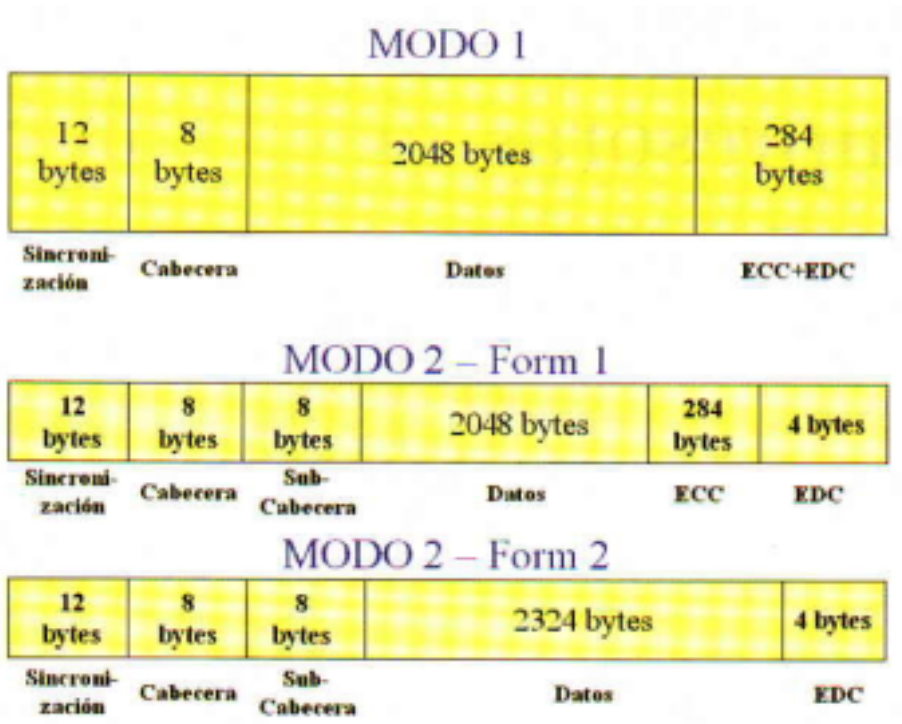


Figura 4. Contenido de los sectores modo 1 y modo 2(incluyendo las variantes form 1 y form 2

La principal diferencia entre un CD-ROM y un CD-audio es el formato de almacenamiento de los datos. En un CD-ROM, el área de datos se divide en sectores, cuyo contenido difiere del presente en las estructuras propias de un CD-audio. Existen dos tipos de sectores: modo 1 (empleados en los CD-ROM) y modo 2 (utilizados en los CD-ROM XA, variante derivada del CD-ROM). Los sectores modo 2 se dividen, a su vez, en dos variantes, denominadas *form1* y *form2*. La Figura 4 muestra la estructura de estos tipos de sectores. Nótese que los sectores modo 1 contienen información para la detección y corrección de errores (EDC+ECC), ya que en un CD-ROM es imprescindible asegurar que cualquier error será corregido.

La variante modo 2 – *form* 1 es muy similar a la variante modo 1, y por tanto sigue siendo adecuada para el almacenamiento de software y datos informáticos. En cambio, la variante *form* 2 sacrifica la información de corrección de errores (ECC), ganando así más espacio para almacenar datos. Esto significa que el CD tendrá mayor capacidad de almacenamiento, pero no será capaz de corregir errores (tan sólo podrá detectarlos). Estos sectores son apropiados

para almacenar información de audio y vídeo. Como ya hemos comentado, con dicho tipo de información se puede consentir una pérdida no excesiva de información.

Los lectores de CD-ROM siguen empleando la técnica CLV, es decir, velocidad lineal constante. En cambio, los lectores más rápidos utilizan CAV. Esto se hace para que la velocidad máxima de giro (correspondiente a la primera vuelta de la espiral) no sea excesiva. El hardware del lector tiene en cuenta que la velocidad de lectura de bits varía, compensando este efecto. Otro aspecto interesante de los lectores de CD-ROM radica en que frecuentemente están preparados para la reproducción de CD-audio, lo que los dota de una doble funcionalidad.

Otra faceta interesante se centra en la posibilidad de almacenar audio y datos en un mismo CD. Para ello, una técnica consiste en almacenar los datos en la primera pista del disco, dejando el resto de pistas para la información de audio. Esto puede dar problemas, puesto que algunos equipos de audio pueden intentar reproducir la primera pista, lo que conduce a errores. Incluso aunque no lo hagan, en algunos casos se puede enviar al láser hacia la pista de datos mediante la tecla de "rebobinado". Para solucionar esto, Philips y Sony desarrollaron el formato CD Extra (también conocido como CD Plus). Este tipo de CD contiene dos sesiones: la primera contiene hasta 98 pistas de audio y la segunda contiene una pista con datos en formato CD-ROM. Un reproductor de audio sólo verá la primera sesión, con lo que se soluciona el problema.

Sistemas de ficheros

Tal y como ocurría con los discos magnéticos, los usuarios de PC no desean trabajar con sectores, sino con estructuras de almacenamiento de todavía mayor nivel de abstracción. Por ello se define un sistema de archivos que hace que el usuario trabaje con ficheros, un concepto mucho más cercano al entendimiento humano.

El formato estándar más empleado en los CD-ROM es el denominado ISO 9660. Éste comienza en el instante 00:02:16, o lo que es lo mismo, en el sector 166 (que es el sector 66 de la pista 1). En los CD-ROM multisesión, el sistema de ficheros ISO 9660 se almacena en la primera pista de datos de cada sesión que contenga datos CD-ROM. La primera especificación de este sistema de ficheros presentaba varias deficiencias, sobre todo relacionadas con el uso bajo Windows 95. Para solventarlas, se creó una extensión de dicho sistema, denominada JOLIET.

A continuación, vamos a repasar las deficiencias corregidas. En primer lugar, los nombres de ficheros sólo podían contener mayúsculas, números y guiones bajos, y estaban limitados a 8 caracteres con 3 caracteres de extensión. Además, la profundidad de subdirectorios estaba limitada a 8 niveles. Finalmente, el formato de nombres de directorios estaba bastante limitado.

Existen otros sistemas de ficheros ampliamente conocidos, como HFS (utilizado en sistemas Macintosh) y ECMA 168 (sistemas UNIX). Estos sistemas quedan fuera del propósito de este artículo, dedicado al mundo del PC, que está ampliamente ligado a los sistemas Windows.

Soportes CD-R y CD-RW

El CD-ROM es un soporte muy adecuado para el almacenamiento de grandes cantidades de datos, pero perdería gran parte de su potencial si los datos sólo pudieran ser grabados en fábrica. Por ello nació la variante CD-R (Compact Disc Recordable o CD grabable). Un CD-R se puede grabar desde un PC, pero una vez los datos se han grabado, ya no es posible borrarlos. Por ello, también se les denomina WORM (*Write Once, Read Multiple*). En cambio, esto no impide que un CD-R se pueda grabar en distintas sesiones. De esta forma, el usuario graba los datos deseados en una sesión, y puede continuar añadiendo datos en futuras sesiones. Lo que no es posible es sobrescribir los datos que ya han sido grabados. El CD-R contiene una espiral pregrabada, pero no está compuesta de aluminio, sino de un pigmento translúcido (recubierto de una capa reflectora). Cuando el láser incide sobre dicha sustancia, ésta se calienta y produce una decoloración. En la lectura, el haz atraviesa la capa translúcida y se refleja en la capa reflectora. En su retorno, la radiación pierde poca intensidad si atraviesa una zona sin decolorar. Si se atraviesa una zona decolorada, se pierde bastante intensidad. De esta forma se simulan los huecos (*pits*) con zonas decoloradas, y las zonas planas (*lands*) con zonas sin quemar. Además, los cambios de intensidad que sufre la luz son muy similares a los que ocurren en un CD grabado en fábrica, por lo que se puede acceder a un CD-R desde un lector de CD-ROM sin problemas.

Para permitir que la información almacenada se pueda borrar y re-escribir, nacieron los CD-RW (*CD Re-Writable* o CD, regrabables). Se basan en las propiedades de cambio de fase de una sustancia, que es la que forma la espiral. En su estado cristalino, dicha sustancia refleja la luz sin problemas. Si se calienta dicha capa hasta una cierta temperatura (mediante el láser), la sustancia pasa a un estado amorfo de baja reflectividad. La propiedad de interés radica en que, si se calienta la sustancia hasta una segunda temperatura (más alta) y se deja enfriar, ésta alcanza el estado cristalino de nuevo. Estas propiedades permiten el re-grabado, puesto que hacen posible el cambio de esta-do en ambos sentidos. Los CD-R y CD-RW son menos tolerantes a las altas temperaturas y a la luz solar que los CD de fábrica. También son más susceptibles a los daños físicos. Además, tras varios (miles) procesos de regrabado, la sustancia que forma la espiral desarrolla cierta tendencia a no cambiar de estado, por lo que es probable la aparición de errores. Si el número de errores no es elevado, los códigos de detección y corrección de errores harán que este efecto sea transparente al usuario. Cuando el número de errores sea excesivo, el disco quedará inservible.

La tecnología DVD

DVD son las siglas de *Digital Versatile Disc*. La tecnología DVD permite fabricar discos ópticos con una gran capacidad de almacenamiento, suficiente para acomodar una película completa en un solo CD. No se debe caer en el error de asociar la tecnología DVD con el cine, lo que recae en el formato DVD-Vídeo.

DVD permite almacenar audio de altísima calidad (mayor que la de un CD-audio) en su formato DVD-audio, e incluso masivas cantidades datos en su formato DVD-ROM.

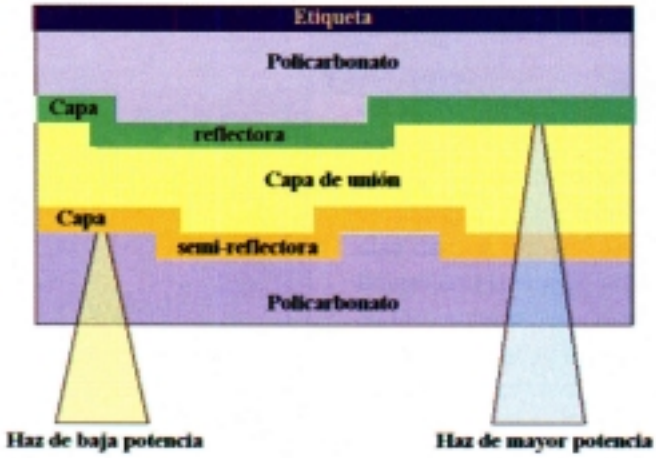


Figura 5. Corte transversal de un disco DVD

Los DVD utilizan la misma filosofía que el CD para almacenar la información, pero permiten emplear una o dos caras, y pueden contener una o dos capas de información en cada cara. La Figura 5 muestra la construcción de un DVD de una cara y dos capas. Cada capa se graba sobre un sustrato de policbonato (como en el CD), y ambos sustratos se enlazan mediante una capa de unión transparente. Además de disponer de hasta dos caras / dos capas, la densidad de información es muy superior a la de un CD. Por todo ello, la capacidad de almacenamiento llega hasta los 17 GB en la variante de 2 caras y 2 capas por cara, pero la fabricación resulta extremadamente compleja, por lo que este formato no es fácil de encontrar.

En caso de disponer de dos capas de información en una cara, la capa exterior es semi-reflectora, mientras que la interior es reflectora. De esta forma es posible acceder a la capa interior, atravesando primero la exterior. Para leer la capa semi-reflectora, el láser se enfoca sobre ésta con baja potencia (ver Figura 5). Si se aumenta la potencia del láser, la capa semi-reflectora es atravesada, y se lee la reflectora. La radiación reflejada atraviesa de nuevo sin problemas la capa semi-reflectora y alcanza al foto-detector. Para finalizar, es interesante anotar que existen equivalentes del CD-RW basados en la tecnología DVD: son los llamados DVD-RAM. Estos alcanzan una capacidad de almacenamiento de hasta 9,6 GB con dos caras de una sola capa.

CAPÍTULO 11

LOS BUSES DEL PC

El número de periféricos estándares que rodean a un PC, así como los muchos periféricos empleados para expandir sus capacidades, hacen que el PC esté formado por un gran número de subsistemas que se comunican entre sí. Llegados a este punto, no cabe duda de que es necesario definir un mecanismo para interconectar dichos bloques funcionales. Ese mecanismo recibe el nombre de bus. En este capítulo vamos a introducir el concepto de bus, y mostraremos la estructura de buses que componen un PC en la actualidad, deteniéndonos en los más conocidos (ISA, PCI y AGP). También introduciremos un componente clave de la placa base, que se halla íntimamente ligado a los buses: el chipset.

El concepto de bus

En primer lugar, un bus es un conjunto de líneas capaces de transportar señales digitales. En la práctica, el concepto de bus se traduce en un conjunto de pistas sobre una placa de circuito impreso (la placa base, en el PC).

Los diferentes dispositivos a interconectar se acoplan al bus, compartiendo las líneas que lo componen. Esto puede llevar a conflictos si no se establece una correcta coordinación de su uso. Por ejemplo, un conflicto importante surgiría si dos dispositivos intentasen poner información sobre el bus al mismo tiempo. Igualmente conflictivo es el caso de un dispositivo que envía información a otro, existiendo un tercer dispositivo que la recibe y procesa (cuando no debería hacerlo). En general, el bus es utilizado por dos dispositivos cada vez: uno envía la información y el otro la recibe. Tras esta discusión, se debe completar aún más la definición de bus con un ingrediente adicional. En efecto, es necesario establecer una serie de normas que aseguren una correcta armonía en el uso del bus (o visto desde el vértice opuesto, que eviten conflictos). En otras palabras, un bus es la suma de un conjunto de líneas digitales (parte física) y un conjunto de normas sobre su uso (parte lógica).

Tras completar la definición de un bus, es importante señalar que, en la situación particular de un bus que interconecta tan sólo dos dispositivos, la denominación correcta sería "puerto" en lugar de bus.

Caracterización de buses

Todo bus se puede caracterizar en función de varios descriptores. Comenzando por su estructura, cualquier bus se puede dividir en dos sub-buses: el bus de direcciones y el bus de datos. El bus de datos es el que transporta la información entre dispositivos durante las transferencias de datos. Por otro lado, el bus de direcciones transporta la información que identifica dónde deben ir a parar los datos transferidos, o bien desde dónde se deben extraer. Dichos

orígenes y destinos se identifican mediante números, comúnmente conocidos como direcciones. Todavía existe un tercer sub-bus, al que se suele denominar bus de control. Éste transporta la información necesaria para controlar el funcionamiento del bus (por ejemplo, una función importante consiste en indicar cuándo hay datos disponibles en el bus).

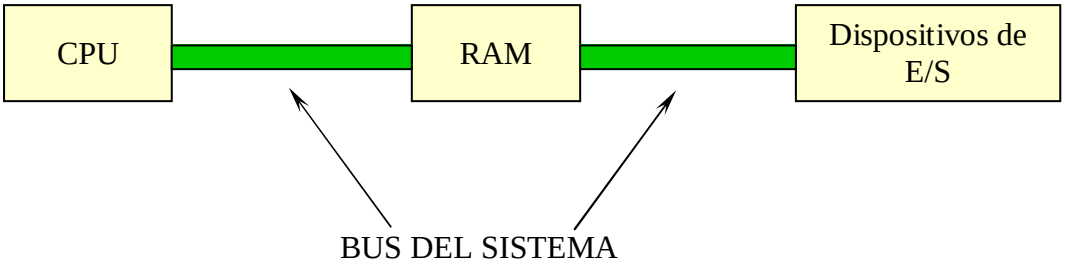


Figura 1. El bus del sistema era el único bus presente en los primeros PC.

Otra característica importante es la anchura del bus. Aunque dicha cantidad está relacionada directamente con el número de líneas que componen el bus, es más frecuente expresarla en bits (en general cada línea transporta un bit de información en cada transferencia). A mayor anchura, el bus será capaz de transferir un mayor número de bits simultáneamente en cada ciclo de bus.

Otro descriptor no menos importante es la velocidad del bus. Ésta refleja la cantidad de ciclos de transferencia que pueden tener lugar por segundo en cada línea del bus, y por tanto se mide en MHz.

De las dos características anteriores se deriva el ancho de banda del bus. Dicha medida se corresponde con la multiplicación de la velocidad y la anchura, y refleja la cantidad de información que puede transferir el bus por unidad de tiempo (en este caso, al tratarse de información, la unidad es MB por segundo). Efectiva mente, cuanto mayor sea el número de líneas disponibles, y mayor sea la velocidad de transferencia en cada línea, mayor será la cantidad total de bytes transferidos por segundo. Tómese como ejemplo el bus ISA, con una anchura de 16 bits y una velocidad de 8 MHz. El ancho de banda se calcularía como sigue: 16 bits x (8 x 1.000.000) ciclos / segundo = 128.000.000 bits / s = 16.000.000 bytes / s = 15,26,MB / s. Nótese la gran diferencia entre MHz y MB / s. Un MHz es 1.000.000 ciclos / s, mientras que un MB / s son 1.048.576 bytes / s. Expresar el resultado del ejemplo como 16 MB / s sería, por tanto, incorrecto por definición.

Jerarquía de buses: el concepto de bus local

En todo PC coexisten varios buses, los cuales se hallan estructurados de forma jerárquica. En primer lugar se encuentra el bus que une a la CPU con la memoria. Este bus se denomina típicamente "bus del sistema". En los primeros PC, los dispositivos de E/S se acoplaban directamente sobre el bus de sistema que, por tanto, era el único bus del PC (ver Figura 1). Esta configuración fuerza a la CPU a transferir datos a la misma velocidad que los dispositivos de E/S, y por tanto se puede pensar en una pérdida de rendimiento del sistema. Sin embar-

go, en aquellos tiempos las CPU trabajaban a velocidades bastante reducidas, por lo que esta configuración era viable.

Hoy en día, forzar a la CPU a trabajar en sincronía con la E/S es impensable. De hecho, no pasaron demasiados años hasta adoptar una solución alternativa: separar a los dispositivos de E/S del bus del sistema. La solución consiste en crear buses dedicados a la E/S (denominados buses locales), y conectar estos al bus del sistema. Los principales buses locales presentes en los PC actuales son: ISA, PCI, AGP y USB. La Figura 2 muestra la jerarquía establecida. Puede apreciarse que al bus ISA se conectan varios dispositivos estándares (especialmente los más lentos), como es el caso de los puertos serie (COM 1 y COM2) y el puerto paralelo (LPT1), además de otros dispositivos de expansión. A1 bus PCI se conectan dispositivos más rápidos, como los discos duros EIDE, los adaptadores de red, y otros muchos dispositivos de expansión a través de las ranuras PCI. Se puede apreciar que AGP es un puerto, ya que tan sólo comunica dos subsistemas (CPU y sistema de vídeo), y por ello no se enlaza al bus del sistema. El bus USB es el más moderno de los tres, y lo explicaremos con detalle en posteriores entregas.

El chipset

En principio, la CPU actúa como un intermediario en las transferencias de información entre dispositivos. En un primer uso del bus, el dispositivo entrega la información a la CPU. En un segundo uso del bus, la CPU envía la información al dispositivo de destino (normalmente, la memoria). Como se puede intuir, esta forma de trabajo proporciona un rendimiento que se puede mejorar. Por ello, los buses suelen aportar técnicas que permiten que los dispositivos intercambien información de forma directa, sin necesidad de pasar por la CPU. Un claro ejemplo es la técnica DMA (*Direct Memory Access* o acceso directo a memoria), que permite que los dispositivos introduzcan o extraigan información en / de la memoria de forma directa. Si no fuera así, para ejecutar un programa habría que volcarlo desde el disco duro a la memoria a través de la CPU, para luego ejecutarlo desde la RAM con un rendimiento adecuado.

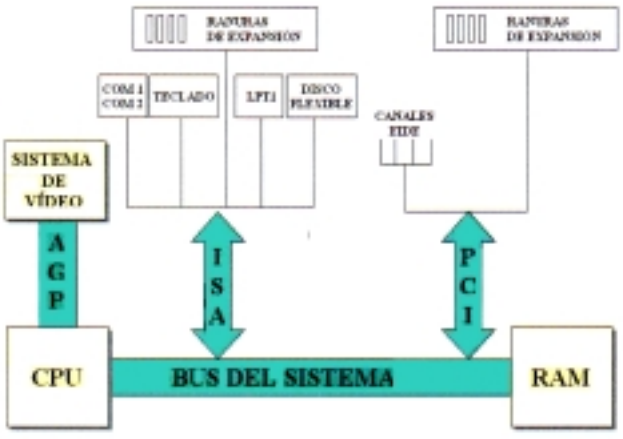


Figura 2. Jerarquía de buses en el PC: los buses locales

Dichas soluciones resultan prácticas desde el punto de vista del rendimiento, pero la complejidad del funcionamiento del bus se ve incrementada. En efecto, los propios dispositivos pueden iniciar las transferencias entre dispositivos. En un momento dado, es fácil que existan varios dispositivos deseando tomar el bus para iniciar una transferencia, lo que podría llevar a conflictos. Esto obliga a implementar un avanzado sistema, de gestión del bus, que determine cuál de los dispositivos puede tomarlo

en cada momento. Dicha función se encuentra implementada en un conjunto de chips que recibe el nombre de chipset.

En realidad, el chipset es mucho más que eso, se trata de un componente imprescindible en cualquier PC, un auténtico protagonista de la placa base. Actúa como un elemento intermediario en las transferencias entre cualquier par de dispositivos del PC. Si un paquete de información fluye hacia la CPU, pasará antes por el chipset. Si la CPU debe ser esquivada, es el chipset quien controla el proceso e implementa la técnica apropiada (por ejemplo, DMA). El chipset también actúa como un puente adaptador entre los diferentes buses (que trabajan de forma muy distinta).

La importancia del chipset es tal, que la E/S y la memoria no podrían trabajar con la CPU sin su existencia. Además, su evolución va ligada a la evolución del PC: las mejoras tecnológicas introducidas en los PC hacen necesaria la introducción de nuevas versiones del chipset. Para ser precisos, la realidad es justamente opuesta: los fabricantes de placas base generan nuevas versiones de forma que puedan acomodar las nuevas versiones de los chipset. Por tanto, la calidad del chipset define la calidad de una placa base. Y teniendo en cuenta que el chipset es un elemento fijo, la única forma de mejorar una placa base es adquirir otra nueva que esté basada en un chipset más avanzado.

El mercado de los chipset está dominado, actualmente, por un reducido conjunto de fabricantes. La mayor parte del mercado la domina Intel, seguido por VIA y SiS. Otros fabricantes, con menor porcentaje de mercado, son ACER y UMC. La Figura 3 muestra el aspecto del chipset Intel 850, que va ligado a las placas base basadas en el procesador Pentium 4, y que proporciona una velocidad de 400 MHz en el bus del sistema.

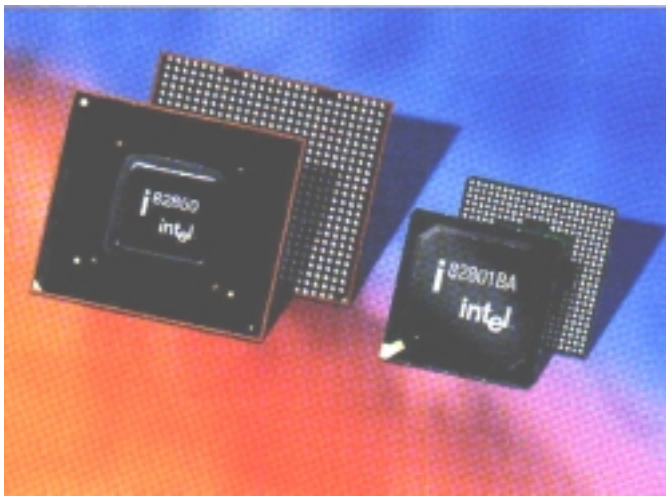


Figura 3. El chipset Intel 850

Una vez introducidos los conceptos generales ligados a los buses del PC, los siguientes apartados recorren con mayor detalle tres de los buses locales más importantes en todo PC: ISA, PCI y AGP.

El bus ISA

Uno de los buses más comunes en el mundo del PC es el bus ISA (*Industry Standard Architecture* o arquitectura estándar industrial). En efecto, se trata de un auténtico estándar ampliamente respetado por los fabricantes de ordenadores PC. Por otro lado, también se trata de un bus que apenas ha cambiado desde su expansión a una anchura de 16 bits, en el año 1984. Dicho de otro modo, este bus ofrece unas prestaciones muy bajas, y totalmente fuera de línea con los requerimientos de los PC de hoy. La razón por la cual persiste es que todavía se sigue fabricando multitud de dispositivos ISA. Además, esto

permite utilizar dispositivos adquiridos para antiguos PC sobre los PC modernos. Finalmente, hay que tener en cuenta que existen muchos dispositivos que no requieren de un bus de altas prestaciones (por ejemplo, los módems estándares), por lo que ISA es más que suficiente en dichos casos.

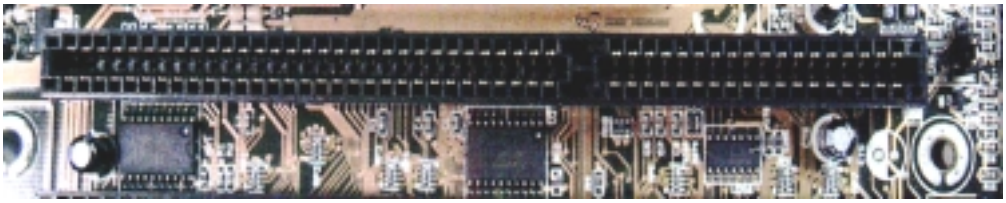


Figura 4. Ranura de expansión ISA

El bus PC original presentaba una anchura de tan sólo 8 bits, totalmente en línea con el primer procesador empleado en el mundo del PC: el Intel 8088 (de 8 bits). Como es de esperar, el bus trabajaba justamente a la misma velocidad que dicho procesador: 4,77 MHz. Con la introducción del procesador 80286 en el IBM AT (en 1984), la anchura del bus se duplicó a 16 bits, y la velocidad ascendía hasta 8 MHz (también totalmente en línea con las características de dicho procesador). Sin embargo, el bus ISA requería de 2 a 3 ciclos para mover 16 bits de datos. Los problemas de compatibilidad antes expuestos hicieron que el bus ISA se estancara en ese último estado, en el que ha llegado hasta nuestros días sin modificaciones. Las ranuras de expansión ISA (Figura 4) dividen sus terminales en dos regiones. Cuando se conectan dispositivos de 8 bits, se emplea tan sólo una región. En el caso de dispositivos de 16 bits, se emplean ambas regiones.

El bus ISA se divide en dos "ramas" (ver Figura 2). Una es interna al PC, y a ella se conectan los dispositivos más lentos del PC: puertos serie, teclado, puerto paralelo, unidad de disco flexible, etc. La otra rama es externa, y se accede a ella a través de las ranuras de expansión ISA.

Además de un rendimiento pobre, otra de las desventajas del bus ISA es la falta de inteligencia en el control del bus. En efecto, todas las transferencias deben ser controladas por la CPU, quedando ésta bloqueada (lo que evita lanzar cualquier otra transferencia). Esto se puede comprobar fácilmente en la práctica desde el sistema operativo: cuando se accede a un disco flexible, se aprecia que el PC queda prácticamente bloqueado hasta que se finaliza la tarea.

Otra desventaja adicional radica en la configuración de los dispositivos, que se debe realizar manualmente. Esto hace posible la aparición de conflictos al instalar nuevas tarjetas.

El bus PCI

PCI (*Peripheral Component Interconnect*) es, con diferencia, el bus local más popular hasta la fecha en el mundo del PC. Fue introducido por Intel en 1993 y, aunque estaba orientado a procesadores Intel de quinta y sexta generación, se utilizaba también en las placas basadas en el procesador 486. El bus PCI presenta normalmente una anchura de 32 bits, y proporciona una velocidad de 33 MHz. Estos datos proporcionan un ancho de banda extremadamente su-perior

al proporcionado por el bus ISA. Además, PCI es el bus de propósito general más rápido que se puede encontrar hoy en día en un PC (para evitar confusiones, es importante decir que AGP no es un bus, sino un puerto, y que no es de propósito general).

Uno de los puntos clave del bus PCI se centra en el chipset asociado. Éste proporciona un control y arbitraje avanzado del bus, lo cual se traduce en un elevado rendimiento; al contrario que ocurría con el bus ISA, el bus PCI permite que los dispositivos inicien transferencias sin la ayuda de la CPU como intermediario. En otras palabras, el bus PCI aporta inteligencia respecto a ISA.

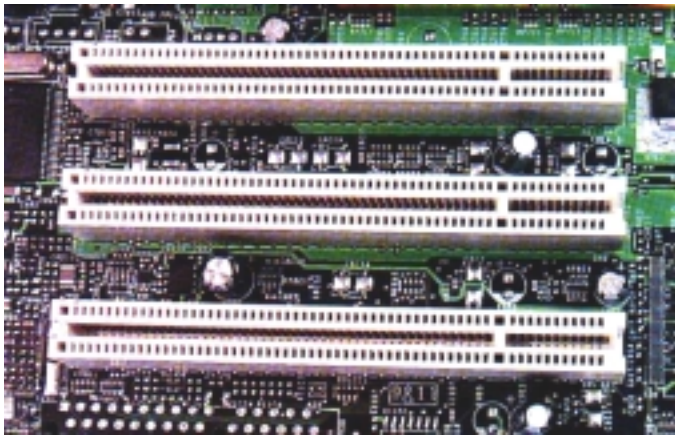


Figura 5. Ranuras de expansión PCI

Otra característica -con importante impacto en el rendimiento- es la posibilidad de transferir datos en modo "ráfaga". Esto permite que, tras proporcionar una dirección inicial, varios conjuntos de datos se transfieran de forma continuada, formando una única transferencia en lugar de muchas.

Para mejorar aún más las prestaciones, las transferencias se realizan a través de zonas de almacenamiento intermedio (*buffers*). De esta forma, la CPU puede enviar los datos al *buffer* y continuar su trabajo, y el bus ya tomará los datos del buffer cuando esté listo para ello (algo análogo ocurre en sentido inverso).

El bus PCI se emplea también fuera del mundo del PC, detalle que lo dota de un cierto grado de universalidad. De hecho, existe una versión del bus PCI dotada de una anchura de 64 bits y una velocidad de 66 MHz (lo cual incrementa el rendimiento en un factor de 4), que se suele emplear en plataformas diferentes al PC (Alpha, etc.).

Y no menos importante es la compatibilidad de PCI con el estándar *Plug and Play* (PnPj lanzado por Intel, en conjunto con multitud de compañías como Microsoft. Esta tecnología se encarga de la configuración automática de los dispositivos, en colaboración con la BIOS y el sistema operativo. Por tanto, ya no es necesario configurar una tarjeta mediante interruptores (*jumpers*). Tan sólo se debe insertar en una ranura de expansión libre, y el sistema se encarga del resto.

El bus PCI también se caracteriza por disponer de dos "ramas". En la rama interna se conecta la controladora de discos, que ofrece hasta 4 canales EIDE sobre la placa base (que se suelen utilizar para conectar unidades de almacenamiento, como discos duros). A la rama externa se accede a través de las ranuras PCI disponibles sobre la placa base, y permite conectar dispositivos externos de expansión.

El número máximo de ranuras de expansión PCI (Figura 5) queda determinado por el chipset que emplea cada placa base. Por ejemplo, una placa basada en el chipset Intel 440BX soporta hasta 5 ranuras PCI, mientras que el *chipset* Intel 815E soporta hasta 6. En general, el número típico es 3/4 ranuras en la mayoría de los PC.

La cantidad de dispositivos PCI disponibles es muy elevada. Como se puede intuir, las grandes prestaciones que ofrece PCI hacen que dichos dispositivos suelan caracterizarse por una elevada velocidad: adaptadores SCSI, tarjetas de red de alta velocidad, discos duros, etc. Como hemos visto antes, los dispositivos más lentos del PC se conectan al bus ISA, ya que no tiene demasiado sentido utilizar con ellos un bus rápido como es PCI.

El puerto AGP

Cada día más, los desarrollos de software para PC requieren de mayores capacidades gráficas. Esto se traduce en la necesidad de incrementar el ancho de banda en las transferencias entre la CPU y el subsistema de vídeo del PC. A pesar de sus elevadas prestaciones, el bus PCI no es apropiado para solucionar este problema, puesto que acomoda demasiados dispositivos, compitiendo por el mismo ancho de banda (y las tarjetas de vídeo requieren un ancho de banda muy elevado, cada vez mayor). En respuesta a dichos requerimientos, surge (en 1997, por Intel) el desarrollo de un bus local de elevadas prestaciones, denominado AGP (*Accelerated Graphics Port*). Como su nombre indica, se trata de un puerto, y por tanto conecta tan sólo dos bloques funcionales: el subsistema de vídeo y la CPU (ver Figura 2). Además, dada su condición de puerto, no es posible expandirlo. De esta forma, la CPU y el sistema de vídeo se comunican a alta velocidad sin interferir con el bus PCI: el ancho de banda que tomaría el sistema de vídeo está ahora disponible para otros dispositivos.

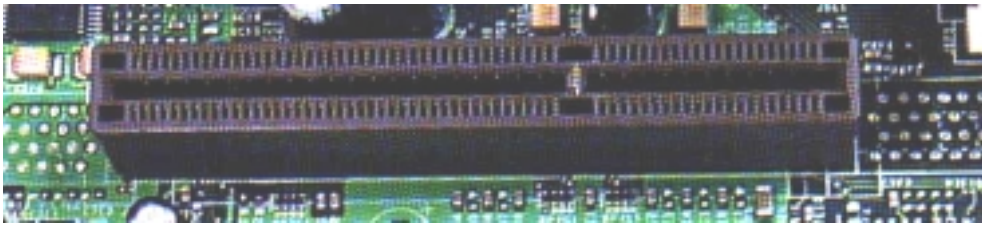


Figura 6. La ranura AGP

El primer soporte para el bus AGP lo proporcionaba el chipset 440LX, diseñado para procesadores Pentium II.

La interfaz que propone AGP es muy similar a la propia de los buses PCI, en muchos aspectos. La ranura AGP empleada por la tarjeta de vídeo (Figura 6) tiene un aspecto similar a las ranuras PCI, pero -dentro de la placa base- se encuentra alejada de estas últimas. De hecho, la especificación AGP se basa en la especificación PCI 2.1 (versión avanzada de PCI antes comentada). Para estrechar aún más la similitud AGP/PCI conviene señalar que, normalmente, las placas base con soporte AGP eliminan una de las ranuras PCI respecto a las placas base sin ranura AGP.

La anchura del bus AGP es de 32 bits, y trabaja a 66 MHz, lo que consigue duplicar el ancho de banda respecto al bus PCI. También existe un modo denominado "2X" que permite duplicar la cantidad de bits transmitidos, utilizando la misma señal de reloj. PCI sólo emplea una de las transiciones de dicha señal (ascendente o descendente), mientras que AGP 2X emplea ambas.

Esto conduce a duplicar el ancho de banda respecto a AGP, y multiplicarlo por 4 respecto a PCI. Aún más, existe la variante "4X", que duplica las prestaciones de AGP 2X.

Entre las características más importantes de AGP, cabe destacar la aplicación de *pipelining* en el acceso a memoria, y la capacidad de compartir la memoria principal con la memoria de la tarjeta de vídeo. Esto último permite emplear mayores cantidades de memoria cuando es necesario (por ejemplo, en operaciones 3D) sin necesidad de instalar memoria sobre la propia tarjeta de vídeo.

CAPÍTULO 12

LA TARJETA DE VÍDEO

A pesar de disponer de una CPU muy potente, un PC no serviría de mucho sin la posibilidad de visualizar la información generada. Cualquier usuario desea obtener la información en forma de texto y/o imágenes; sólo así será fácil de interpretar por un ser humano. En un extremo se encuentra la CPU, que puede generar información visual (texto y gráficos), pero siempre representada en forma digital (unos y ceros). En el otro extremo, el monitor, un componente controlado mediante señales analógicas (digitales en el pasado), que indican qué puntos de la pantalla hay que iluminar, y con qué características (color, etc.). Queda clara, por tanto, la necesidad de un tercer elemento que actúe como intermediario o interfaz. Dicho componente se encarga de tomar la información digital que ofrece la CPU, y generar las señales apropiadas para controlar el monitor, haciendo que la información sea visible para el usuario. Tal subsistema se materializa en forma de un periférico estándar, conocido como tarjeta de vídeo (la Figura 1 muestra el aspecto de una tarjeta de vídeo actual). Los siguientes apartados se adentran en el mundo de las tarjetas de vídeo, mostrando su funcionamiento y características fundamentales. Como apreciará, dichas tarjetas son, actualmente, mucho más que un mero intermediario.

Figura 1. Aspecto de una tarjeta de video (basada en el chip GeForce 2 de nVidia)

Funcionamiento de la tarjeta de vídeo

Las primeras tarjetas de vídeo ligadas al mundo del PC eran realmente sencillas. Tan sólo se encargaban de tomar la información producida por la CPU, y crear la imagen correspondiente sobre el monitor del PC. En otras palabras, funcionaban prácticamente como un convertidor digital/analógico. Esto era llevadero, puesto que las imágenes a visualizar se componían simplemente de texto, y ni siquiera se aplicaba color. Por tanto, las imágenes a mostrar se componían de una reducida cantidad de datos.

Con el paso del tiempo, los sistemas operativos gráficos (principalmente, Windows) entraron en escena. El PC se ha visto obligado a manejar un volumen de información extremadamente mayor: imágenes con una gran resolución y miles (o millones) de colores. Si la CPU tuviera que producir tal cantidad de información y, aún más, operar con ella, el tiempo de cálculo necesario sería prohibitivo, y tendría un impacto negativo en el rendimiento del sistema. Para hacerse

una idea, imagine el "esfuerzo" que supondría para la CPU tener que dibujar entidades como ventanas, cursores e iconos, y mover (re-dibujar) toda esa información conforme sea requerido. La CPU invertiría un altísimo porcentaje de su tiempo en operaciones gráficas, evitando que los programas se ejecutasen de forma ágil.

Por ello, las tarjetas de vídeo se convirtieron en algo más que un intermediario entre la CPU y el monitor, y pasaron a denominarse "tarjetas aceleradoras". Además, la popularidad de Windows en el mundo del PC hizo que dichas tarjetas se conocieran comúnmente como "tarjetas aceleradoras de Windows". La mejora fundamental consiste en añadir inteligencia a la tarjeta, es decir, capacidad de procesamiento. Usando una tarjeta aceleradora, cuando el sistema debe dibujar una ventana, la CPU no debe generarla punto a punto. Por el contrario, la CPU simplemente indica a la tarjeta que quiere dibujar dicha ventana con ciertas características, y la tarjeta se encarga de hacer las operaciones necesarias para hacer efectivo el trazado en pantalla. De esta forma, la CPU no se dedica a las operaciones gráficas, y emplea su tiempo para otras tareas, con un importante incremento del rendimiento del PC. Por su parte, la tarjeta aceleradora realiza su trabajo a mayor velocidad que lo haría la CPU, ya que está optimizada para ello. De todo lo anterior se deduce que la tarjeta aceleradora se comporta como un co-procesador destinado a operaciones gráficas (tal y como el co-procesador matemático hacía con las operaciones en coma flotante).

Teniendo en cuenta la estrecha relación entre el PC y los sistemas operativos gráficos, podemos intuir que hoy cualquier tarjeta de vídeo cumple con la función aceleradora.

Sobre todo, se debe retener una idea principal: las tarjetas de vídeo son, actualmente, un componente crítico. Si no son lo suficientemente inteligentes, rápidas, dotadas de memoria y ágilmente comunicadas con la CPU, el rendimiento del PC se verá afectado en gran medida.

Componentes de una tarjeta de vídeo

Tal y como se muestra en la Figura 2, la tarjeta de vídeo se compone de tres subsistemas fundamentales: el vídeo chipset, la memoria de vídeo (conocida en inglés como *frame buffer* y el RAMDAC.

Figura 2. Estructura y funcionamiento de una tarjeta de vídeo.

La función aceleradora de las tarjetas implica capacidad de cálculo, y, por tanto, exige la existencia de un procesador integrado en la tarjeta. Hace años, dicho procesador se implementaba mediante un conjunto de chips, y de ahí el nombre de vídeo chipset. Actualmente, las tecnologías de integración permiten implementar tal procesador en un único chip o GPU (en la Figura 1, cubierto por un ventilador), que no es raro que se carac-

terice por una complejidad superior a la de algunas CPU como el Pentium III. Muchos fabricantes diseñan sus propios chipsets para sus tarjetas de vídeo, como es el caso de Matrox. Esto permite un gran control sobre el diseño y el desarrollo de controladores eficientes. Esta aproximación implica un mayor esfuerzo en el desarrollo de la tarjeta pero, con frecuencia, da como resultado un producto muy optimizado.

En cambio, otros fabricantes toman chipsets de terceros y los integran en el diseño de su tarjeta. Por ejemplo, la firma Diamond Multimedia fabricaba la tarjeta Diamond Monster 3D II, que está basada en el chipset Voodoo 2 (de la firma 3Dfx). Entre los más conocidos fabricantes de chipsets se encuentran Intel, ATi, Matrox y nVidia. En la Figura 4 se puede apreciar una de las GPU más populares, el GeForce de la firma nVidia.

Tras las operaciones gráficas que realiza el chipset, el resultado es la información de vídeo a mostrar en el monitor, siempre en formato digital. Al igual que la CPU necesita a la RAM para almacenar los resultados de las operaciones, la GPU también necesita almacenar la información de vídeo resultante en una memoria. En los inicios del PC, el requerimiento de espacio no era mucho (ya que se trabajaba básicamente con texto). De hecho, bastaba con unos 2 kB para almacenar una pantalla de texto monocromo. Por ello, se empleaba una zona especial de la memoria superior del sistema (UMA) para alojar los datos de vídeo. La CPU colocaba allí la información a visualizar, y la tarjeta de vídeo la tomaba de allí para enviarla al monitor. Hoy en día, la cantidad de datos a manejar es tan grande que no resultaría nada beneficioso emplear la memoria del sistema para tareas de visualización. Por ello, las tarjetas de vídeo integran su propia memoria RAM. La cantidad y el rendimiento de la memoria empleada tienen tanta importancia como la eficiencia del chipset. Por ello, es importante prestar atención a las especificaciones del fabricante: cuánta memoria se integra en la tarjeta y cuáles son sus prestaciones.

Figura 3. El chip ATI Mobility Radeon 7500

Por el momento, la información producida por el chipset (y almacenada en la memoria) se encuentra en formato digital. Por ello hace falta todavía un elemento que convierta dicha información en señales analógicas para gobernar el monitor. Dicho componente recibe el nombre de RAMDAC (RAM porque toma la información de una memoria RAM, y DAC de Digital to Analog Converter, conversor digital /analógico). Volviendo hacia atrás en el tiempo, se puede apreciar que las antiguas tarjetas de vídeo se comportaban prácticamente como un RAMDAC. La operación de leer la información digital y producir imágenes visibles se realiza multitud de veces por segundo. Como era de esperar, las prestaciones del RAMDAC tienen un gran impacto en la calidad de la imagen, una vez en panta-

lla. En concreto, el RAMDAC condiciona la frecuencia de refresco de la pantalla, la máxima resolución obtenible, y el máximo número de colores que se pueden visualizar.

Una vez vistos los tres componentes de una tarjeta de vídeo, es interesante citar su aparición en un producto real. Por ejemplo, la tarjeta Millennium G550 (de la firma Matrox) integra el chip G550 (de la misma firma), además de 32 MB de RAM y un RAMDAC a 360 MHz.

La tarjeta de vídeo y los buses del PC

Como ya hemos comentado, la tarjeta de vídeo es un componente crítico con respecto al rendimiento del PC. Cuanto mejores prestaciones tenga la tarjeta, mejor responderá el PC en términos generales. Estos requerimientos se trasladan directamente a los componentes internos antes presentados: las prestaciones del chipset, la memoria y el RAMDAC son las que condicionan el rendimiento general de la tarjeta de vídeo. Siendo más precisos, la afirmación anterior sólo es cierta si se consideran los tres componentes en conjunto. En efecto, de poco sirve tener un chipset muy rápido si la memoria o el RAMDAC son lentos. En general, el componente con menores prestaciones limita a los otros dos, así que es necesario que los tres componentes se diseñen de forma que el rendimiento combinado sea óptimo.

Pero, lamentablemente, no todo acaba en la tarjeta: debe existir una comunicación fluida entre la tarjeta de vídeo y la CPU. De hecho, la tarjeta de vídeo requiere más ancho de banda que cualquier otro periférico. Por tanto, las características del bus local empleado tienen una gran influencia sobre las prestaciones finales.

Figura 4. El chip GeForce 3 de nVidia

A lo largo de la historia del PC han existido tarjetas de vídeo para todo tipo de buses. Pero más bien habría que decir que los buses han ido evolucionando con las tarjetas de vídeo, ya que éstas han ido incrementando sus requerimientos de ancho de banda sin cesar.

En primer lugar, se encuentran las tarjetas para el bus ISA. Como se puede intuir, se caracterizan por sus bajas prestaciones, y no están a la altura de los requerimientos actuales. Sin embargo, todavía tienen utilidad como elemento de test. Si la tarjeta de vídeo causa conflictos con otros dispositivos, y el sistema no arranca, siempre es posible sustituir la tarjeta en uso por una tarjeta ISA. De esta forma se podrá detectar si el fallo se debe a la tarjeta de vídeo empleada o lo causa otro dispositivo.

Con la llegada del procesador 486, el bus local VESA (también conocido como VLB) alcanzó una notable popularidad. Dicho bus es fácil de reconocer por los conectores empleados, cuyo tamaño supera al de los conectores ISA. El bus

VLB estaba más cerca del bus PCI que del bus ISA. Por tanto presentaba unas muy buenas prestaciones, muy superiores a las obtenidas con el bus ISA. Es importante remarcar que VESA es una asociación de fabricantes relacionados con el hardware de vídeo lo cual remarca cómo las tarjetas de vídeo han condicionado la evolución de los buses del PC. El VLB está limitado a los PC basados en el procesado: 486, aunque también es posible encontrarlo en los sistemas Pentium más antiguos.

Con la llegada del procesador Pentium, el bus PCI tomó el relevo al bus VLB, extendiendo su popularidad hasta la actualidad. Las tarjetas de vídeo PCI mejoran las prestaciones de las tarjetas VLB, alcanzando las máximas prestaciones jamás obtenidas en el PC. Además, el bus PCI introduce nuevos conceptos como Plug and Play (del que hablamos en el anterior capítulo).

El bus PCI ha sido mejorado en la actualidad, con la introducción del bus AGP. Como ya se introdujo, se trata realmente de un puerto (y no de un bus), que une exclusivamente la tarjeta de vídeo con el procesador. El bus AGP llega a cuadruplicar el ancho de banda obtenido con el bus PCI, y por ello se encuentra presente en la mayoría de los PC actuales. Nótese, una vez más, cómo los exigentes requerimientos de las tarjetas de vídeo han condicionado el nacimiento de un nuevo bus, esta vez exclusivamente dedicado al hardware de vídeo.

Modos de vídeo, resolución y color

La información de vídeo se caracteriza por varios parámetros, que determinan hasta dónde puede llegar cada tarjeta de vídeo. En primer lugar, hay que distinguir entre dos modos de vídeo fundamentales: el modo texto y los modos gráficos.

En modo texto, la tarjeta de vídeo almacena la forma (definición píxel a píxel) de todos los caracteres que es posible mostrar (generalmente, este conjunto de caracteres se ajusta al estándar ASCII). No es posible acceder a los píxeles que forman cada carácter. La CPU dice a la tarjeta de vídeo qué caracteres quiere mostrar, y la tarjeta de vídeo es la encargada de generar- su forma en pantalla.

En los modos gráficos, por el contrario, los píxeles se manejan de forma individual, y, por tanto, es posible generar tanto texto como imágenes. Al trabajar al nivel de píxel, la cantidad de información a almacenar para llenar una pantalla es extremadamente superior.

Actualmente, cualquier tarjeta es capaz de mostrar información tanto en modo texto como en modos gráficos. En cambio, las primeras tarjetas empleadas en el mundo del PC trabajaban exclusivamente en modo texto.

Otra distinción a realizar se centra en los modos monocromo y color. El modo monocromo, en el ámbito práctico, no existe hoy en día en el mundo del PC (salvo en muy contadas aplicaciones específicas). En modo color, cada píxel se identifica mediante un grupo de bits, que determinan el color a aplicar. Realmente, el color se crea como la suma de tres colores primarios: rojo, verde y azul (RGB). Por ello, el conjunto de bits se divide en tres grupos, que identifican

la intensidad a aplicar a cada color primario. El número de bits empleados por píxel determina la cantidad de colores que es posible mostrar. En el modo VGA estándar se emplean 4 bits por píxel, lo que da lugar a 16 posibles colores. Empleando 8 bits se obtiene el modo de 256 colores.

El modo high color obtiene 65.535 colores, puesto que aplica 16 bits. Finalmente, existe un modo denominado true color (color verdadero) que emplea un byte para cada color primario (24 bits por píxel). Se obtienen más de 16 millones de posibles colores, lo que da lugar a imágenes de una calidad extremadamente alta, adecuada para aplicaciones de edición fotográfica, diseño gráfico, etc. Al emplear 24 bits por píxel, la cantidad de memoria necesaria para almacenar cada pantalla es muy elevada. Por ello, fuera de las aplicaciones profesionales antes citadas, es conveniente utilizar el modo high color, puesto que requiere un menor uso de la memoria de vídeo.

Un parámetro fundamental es la cantidad de píxeles que componen una pantalla generada por la tarjeta de vídeo, dato conocido como resolución. La resolución se suele especificar con el formato "H x V", donde H es el número de píxeles horizontales y V es el número de píxeles verticales. Las resoluciones más comúnmente utilizadas en los PC quedan recogidas en la Tabla 1. También se indica el número de píxeles por pantalla para cada resolución (la multiplicación de H por V).

No hay que confundir dichas resoluciones (las que la tarjeta de vídeo puede generar), con la resolución máxima que puede ofrecer el monitor. Por ejemplo, suponga que dispone de una tarjeta capaz de trabajar con una resolución de 1.600 x 1.200. No podrá emplear dicha resolución tan elevada con un monitor que sólo es capaz de ofrecer hasta 1.024 x 768.

RESOLUCIÓN	NÚMERO DE PÍXELES POR PANTALLA
320 x 200	64.000
640 x 480	307.200
800 x 600	480.000
1.024 x 768	786.432
1.280 x 1024	1.310.720
1.600 x 1200	1.920.000

Tabla 1. Resoluciones de uso común en el mundo del PC.

Llegados a este punto, queda patente que la combinación de número de colores y resolución permite definir varias configuraciones para la tarjeta de vídeo, que requerirán mayor o menor memoria para el almacenamiento. Por ejemplo, trabajando con una resolución de 800 x 600 y high-color, se tendrán: (800 x 600) píxeles x 16 bits/píxel = 0,92 MB por pantalla.

No todas las tarjetas de vídeo aceptan todas las posibles combinaciones. Esto se puede apreciar en el panel de control de Windows (configuración de pantalla). La resolución y número de colores que Windows permite seleccionar cambia de un PC a otro, en función de las características de la tarjeta de vídeo instalada.

Frecuencia de refresco

Cada segundo, el monitor del PC se llena de contenido gráfico (píxeles) un cierto número de veces. Dicho número se conoce como frecuencia de refresco, se mide en hercios (Hz) y es otro importante parámetro que caracteriza a las tarjetas de vídeo. El RAMDAC acude a la memoria de vídeo para tomar los píxeles a mostrar, convirtiendo dicha información en señales analógicas que se envían monitor. Por tanto, la frecuencia de refresco máxima queda limitada por la rapidez con la que el RAMDAC puede realizar su trabajo (la velocidad del RAMDAC se mide en MHz, indicando la frecuencia con la que éste procesa los píxeles y envía las señales de vídeo al monitor).

Las frecuencias de refresco se encuentran estandarizadas, para mejorar la compatibilidad entre tarjetas de vídeo y monitores de distintos fabricantes. Los valores más comunes son: 56, 60, 65, 70, 72, 75, 80, 85, 90, 95, 100, 110 y 120 Hz.

Pero ¿qué importancia tiene dicho parámetro para el usuario? Una frecuencia de refresco apropiada mejora la visibilidad de la imagen y reduce la fatiga ocular. Si la frecuencia de refresco es muy baja, el ojo percibirá que la imagen vibra.

Esto se puede explicar a través de las características del ojo. Cada vez que la retina detecta una imagen, ésta persiste un cierto tiempo hasta poder detectar otra. Si la frecuencia de refresco es menor que la frecuencia de muestreo del ojo, este último podría capturar la imagen en pleno refresco de la pantalla, y de ahí ese molesto efecto de vibración.

Hay que tener en cuenta que la frecuencia de muestreo del ojo varía de un individuo a otro. Por tanto, es interesante probar varias de las frecuencias que la tarjeta de vídeo sea capaz de ofrecer. Basta con seleccionar una frecuencia que elimine el efecto no deseado (en otras palabras, que no fatigue la vista del usuario) aunque no se trate de una frecuencia elevada. En general a partir de 72 Hz los parpadeos desaparecen. Nótese también que las frecuencias muy altas apenas tienen impacto, ya que son muy superiores a las frecuencias de muestreo de cualquier ojo humano, y, por tanto, no notaremos diferencia alguna.

La frecuencia de refresco está íntimamente ligada a la resolución de la imagen. Aumentar la resolución implica mayor cantidad de píxeles a mostrar por pantalla, lo que obliga a aumentar la velocidad del RAMDAC para mantener la misma frecuencia de refresco. Visto de otro modo y con un ejemplo, imagine que trabaja con una resolución de 800 x 600 a 120 Hz. Si aumentara la resolución a 1.600 x 1.200, el número de píxeles por pantalla sería cuatro veces mayor. Para seguir teniendo 120 Hz, el RAMDAC deberá trabajar 4 veces más rápido, y sólo será posible si la tarjeta de vídeo dispone de un RAMDAC capaz de ello.

Estándares de vídeo

Los parámetros vistos hasta ahora permiten definir infinidad de modos de trabajo respecto al hardware de vídeo combinando resoluciones, números de colores, frecuencias de refresco, etc. Si cada fabricante aplicara cualquier valor a

dichos parámetros, la compatibilidad entre tarjetas de vídeo, monitores y software de distintos fabricantes sería más que difícil.

Por ello nacieron los estándares de video cuyo objetivo consiste en estandarizar dichos parámetros para asegurar la compatibilidad. En orden cronológico, estos han sido: MDA (texto en monocromo), Hercules (texto y gráficos en monocromo), CGA (permite 16 colores con una resolución de 160 x 200), EGA (mejora la resolución de CGA), VGA y SVGA. Hay que decir que VGA fue el último estándar bien definido y universalmente aceptado. El estándar SVGA se presenta borroso, y hace referencia a cualquier mejora de VGA, aunque cada usuario y fabricante lo interpreta de una forma distinta. Por fortuna, nació el estándar VESA Super VGA, con la intención de crear orden y claridad en la definición de los modos de vídeo actuales.

CAPÍTULO 13

LA TARJETA DE SONIDO

En la anterior entrega tratamos un componente relacionado con la salida del PC: la tarjeta de vídeo. Dicho componente funcionaba como un coprocesador especializado en vídeo, que ayudaba a la CPU en la tarea de ofrecer información visible. En esta entrega enfocaremos otro tipo de información de salida: la audible. Desde sus comienzos hasta la actualidad, el PC ha podido generar sonidos de baja calidad mediante su altavoz interno. Sin embargo, no es posible escuchar sonido de alta calidad (por ejemplo, temas musicales con calidad de CD) utilizando el altavoz interno. En ese caso es necesario añadir hardware adicional dedicado a la información de audio. Dicho hardware se conoce comúnmente como tarjeta de sonido y, hoy en día, es un elemento clave en cualquier PC (Figura 1). Hay que remarcar que las tarjetas de sonido no suelen ser elementos dedicados exclusivamente a la salida de información. También son capaces de capturar información audible desde el exterior del PC, ya sea mediante un micrófono o cualquier otra fuente de audio. En línea con lo que ocurría con las tarjetas de vídeo, la tarjeta de sonido debe considerarse como un coprocesador dedicado a trabajar con información de audio, liberando a la CPU de dicha carga.

Figura 1. Aspecto de una tarjeta de sonido (SoundBlaster AWE64)

Como ya hemos introducido, las tarjetas de sonido se han convertido en un componente imprescindible en cualquier PC. De hecho, algunos PC se fabrican con el hardware de audio integrado en la placa base (algo que también ocurre con las tarjetas de vídeo). El usuario actual del PC reproduce con frecuencia música en diversos formatos, disfruta de películas en formato DVD (que, por supuesto, incluyen sonido de alta calidad), captura sonido del exterior del PC, asigna sonidos a los eventos lanzados por Windows, etc. Todas estas tareas, y muchas más, serían imposibles sin el uso de la tarjeta de sonido.

Funciones básicas

La mayoría de tarjetas de sonido implementan cuatro funciones básicas: reproducción, captura, síntesis y procesamiento de sonido.

En primer lugar, la tarjeta debe ser capaz de reproducir audio, ya sea desde lectores de CD o DVD, o desde ficheros almacenados en el disco duro, usando formatos estándares como WAV, MP3 y MIDI.

Además, la tarjeta debe ser capaz de realizar el proceso inverso, es decir, almacenar audio procedente de. una fuente externa. Esto incluye capturar sonidos mediante un micrófono, o introducir sonido desde cualquier otra fuente (instrumentos musicales, reproductores de cintas, etc.). La información queda almacenada, generalmente, en el disco duro del PC en forma de ficheros.

La tercera función básica se centra en la síntesis de audio, o lo que es lo mismo, la creación de sonido. Nótese que las dos funciones anteriores se centran -básicamente- en una pura conversión de información entre los mundos analógico (exterior del PC) y digital (interior del PC). La síntesis de audio exige capacidad de procesamiento a la tarjeta.

Finalmente, otra importante función es el procesamiento de sonidos existentes (generalmente almacenados en el disco duro como archivos). De nuevo, la tarjeta de sonido aplica su capacidad de procesamiento, ahorrando todo ese trabajo a la CPU.

Si estas funciones fueran realizadas por la CPU, el rendimiento del sistema se vería afectado negativamente. En la anterior entrega apreciábamos la elevada cantidad de información que implica el procesamiento de datos de vídeo. En el caso del sonido, el volumen de información es menor, pero no deja de ser elevado, y por tanto el papel que desempeña la tarjeta es crucial.

En el ámbito práctico, las funcionalidades antes comentadas hacen posibles tareas como escuchar un CD-audio mientras se trabaja con el PC, escuchar el audio que acompaña a páginas Web, reproducir temas musicales en formato MP3, escuchar los sonidos que acompañan a secuencias de vídeo para PC, disfrutar del chat con voz, crear música con el PC, escuchar los efectos sonoros que acompañan a los juegos, conectar instrumentos musicales al PC, etc.

Es conveniente citar que las tarjetas de sonido se dividen en dos tipos: half duplex y full duplex. Las tarjetas full duplex son capaces de producir (operación de salida) y capturar (operación de entrada) señales de audio de forma simultánea. En cambio, las tarjetas half duplex sólo pueden realizar una de estas operaciones cada vez. Muchas aplicaciones exigen una tarjeta full duplex para su correcto funcionamiento (por ejemplo, aplicaciones de videoconferencia y algunos juegos). Resulta sencillo comprobar esta característica desde Windows. Tan sólo es necesario abrir dos instancias de la grabadora de sonidos. En una de ellas se inicia la reproducción de un fichero de audio, mientras que en la otra se inicia la grabación de sonido. Si el proceso de grabación se puede realizar con éxito, se deduce que la tarjeta de sonido es full duplex.

Componentes fundamentales

El "corazón" de cualquier tarjeta de sonido está formado por tres subsistemas (ver Figura 2): el convertidor analógico/digital (CAD), el procesador digital de

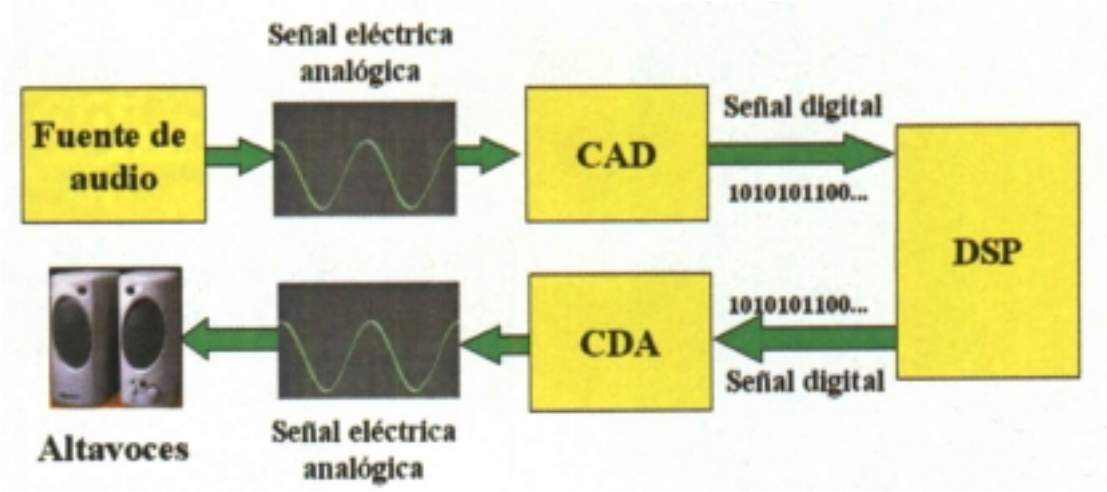


Figura 2. Componentes fundamentales de una tarjeta de sonido.

señales (DSP, *Digital Signal Processor*) y el convertidor digital/analógico (CDA).

El CAD actúa como interfaz con el mundo exterior del PC para la entrada de audio. En el exterior del PC, el sonido se encuentra en forma de ondas de presión (variaciones de presión en el espacio y en el tiempo). Gracias a un transductor primario, que suele ser un micrófono, dichas ondas de presión se convierten en señales eléctricas (variaciones de tensión en el tiempo). El CAD es capaz de tomar muestras de tensión sobre dichas señales, y asignar valores digitales (números binarios) a cada muestra. Con ello, se consigue una representación del sonido en formato digital que, ahora sí, puede ser tratado por un procesador. El uso de un micrófono no es imprescindible, ya que cualquier fuente de audio como un reproductor de cintas o un receptor de radio entrega el sonido -directamente- en forma de señales eléctricas.

El CDA realiza el proceso inverso, implementando la interfaz entre el PC y el mundo exterior para la salida de audio. El CDA toma una secuencia de datos digitales, y transforma dichos datos en niveles de tensión, creando una señal eléctrica analógica. Empleando después un transductor (típicamente unos altavoces o cascos), las señales eléctricas se convierten en ondas de presión, audibles por el ser humano.

Entre el CAD y el CDA se encuentra el DSP, que dota a la tarjeta de capacidad de procesamiento. El DSP es un procesador especializado en el tratamiento de señales digitales, para lo cual es necesaria una elevada capacidad de cálculo, que no es posible obtener mediante procesadores convencionales. Sus características son similares a las de un procesador de propósito general pero, en cambio, su arquitectura es diferente (típicamente Hardware. Además, la organización de la memoria es también diferente. La principal diferencia se centra en la unidad aritmética, que ofrece procesamiento paralelo e incluye unidades especializadas (multiplicadores, etc.). El resultado general es una velocidad de

trabajo de 2 a 3 veces mayor (1os DSP son capaces de realizar millones de operaciones en coma flotante cada segundo). En conclusión, el DSP de la tarjeta de sonido es el centro de tratamiento de audio del PC (por ejemplo, la aplicación de efectos como los ecos se deben a dicho componente). Como se puede intuir, la tarjeta de sonido suele incorporar memoria, como complemento indispensable para el DSP.

Entrando en el campo de la producción de tarjetas de sonido, ocurre algo similar a lo que se mostró para las tarjetas de vídeo. Muchos fabricantes implementan los anteriores elementos mediante sus propios chipsets, mientras que otros toman chipsets de terceros, y le añaden funcionalidad adicional, que caracteriza a su producto.

Elementos de interfaz

Los tres componentes presentados constituyen el núcleo de la tarjeta, pero es necesario complementarlos con ciertos elementos de interfaz. En primer lugar, se encuentra un conjunto de conectores de entrada y salida, que constituyen la interfaz analógica con el mundo exterior, desde el punto de vista del usuario. Uno de los conectores de entrada está preparado para la conexión de un micrófono. Otro conector, comúnmente denominado "entrada de línea", permite introducir señales procedentes de otras fuentes de audio (reproductores de cinta, receptores de radio, etc.). Finalmente, un tercer conector proporciona la salida de audio. A éste se suelen conectar altavoces de sobremesa, auriculares, grabadoras de cintas o cualquier otro tipo de equipo capaz de trabajar con señales de audio analógicas. Las entradas y salidas se conectan a un dispositivo mezclador, que combina las diferentes señales de audio, con niveles individuales controlables por el usuario.

Algunas tarjetas proporcionan entradas y salidas digitales. Éstas permiten introducir la información de audio directamente en formato digital, evitando el CAD y el CDA. Con ello se consigue una mayor calidad en los resultados, ya que los procesos de conversión entre los mundos analógico y digital conllevan una inevitable pérdida de calidad. Por ejemplo, se pueden conectar reproductores de CD a las entradas digitales, y grabadoras DAT o CD-R a las salidas digitales. Es interesante remarcar que existen tarjetas de sonido puramente digitales, que contienen únicamente interfaz para la conexión de equipos digitales.

Por otro lado, muchas tarjetas están dotadas de una interfaz MIDI, que permite conectar instrumentos musicales al PC a través de una interfaz digital estándar. También es frecuente encontrar un puerto de juegos, que permite conectar dispositivos de control como *joysticks* o *gamepads* (componentes esenciales en el mundo de los juegos para PC).

Finalmente, hay que hablar de una interfaz imprescindible: la que comunica la tarjeta de sonido con el bus del PC.

Dicha interfaz caracteriza en gran medida las prestaciones de la tarjeta. En efecto, las tarjetas de sonido ISA suelen ofrecer menores prestaciones que las PCI. Estas últimas son las más comunes -con gran diferencia- en la actualidad. Incluso muchos PC incorporan la tarjeta de sonido en la propia placa base, en

forma de chipset (con lo que se gana una ranura PCI libre para otra tarjeta de expansión).

Muestreo y cuantización

La función de adquisición de señales de audio se basa en dos procesos fundamentales, denominados muestreo y cuantización, íntimamente relacionados con el CAD. No es el objetivo de este apartado mostrar en detalle la teoría que rodea a dichos procesos, pero sí ofrecer un conocimiento general, ya que son fundamentales en la operación de captura de audio.

En primera instancia, se parte de una señal eléctrica analógica, que contiene la información de audio que se pretende digitalizar. Dicha señal, al ser analógica, es continua en el dominio del tiempo y también en el dominio de la amplitud (niveles de tensión). Al ser continua en el tiempo, la señal contiene valores de tensión eléctrica para todos los posibles instantes de tiempo (no hay que olvidar que cualquier intervalo de tiempo contiene un número infinito de instantes de tiempo). Al ser continua en amplitud, la señal puede tomar cualquier nivel de tensión posible (infinitos posibles valores en cualquier rango).

El proceso de muestreo se encarga de "discretizar" el dominio del tiempo. Como su propio nombre indica, dicho proceso consiste en tomar muestras de la señal analógica en distintos instantes de tiempo. Las muestras se suelen tomar en intervalos de tiempo regulares. Tras el muestreo (Figura 3), se obtiene una señal discreta en el tiempo (ya no hay valores para todos los posibles instantes), pero que sigue siendo continua en amplitud (la señal puede tomar cualquier nivel de tensión).

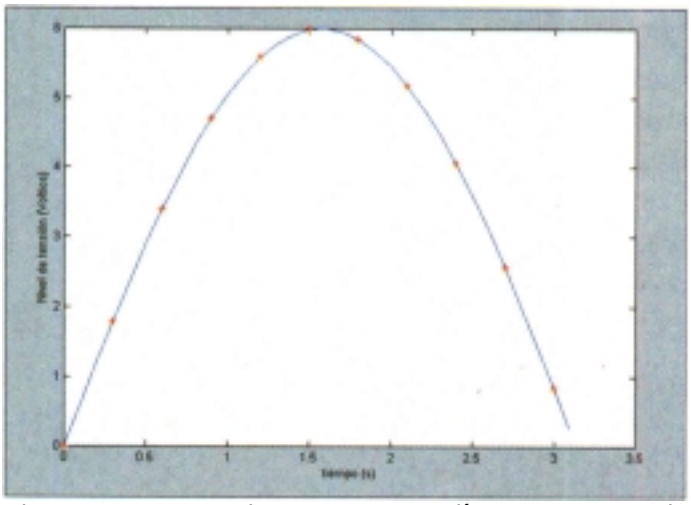


Figura 4. Proceso de muestreo. La línea representa la señal analógica a digitalizar. Los puntos representan las muestras, tomadas cada 0,3 seg. (es decir, con una frecuencia de muestreo de 1/0,3 Hz)

A continuación, se procede a discretizar el dominio de la amplitud. Para ello, se establece un determinado número de niveles de tensión, denominados "niveles de cuantización". Para cada muestra de la señal, el nivel de tensión asociado se aproxima a uno de los niveles de cuantización. Por tanto, la señal resultante sólo puede contener un conjunto finito de niveles de tensión. Cada nivel de cuantización tiene asociado un número binario. Por ejemplo, si se trabajara con 3 bits (ver Figura 4), se podrían establecer 23 = 8 niveles de cuantización, a los que se asignarían los valores binarios 000, 00l, 0l0; 0ll,..., 110 y 111.

El proceso de muestreo queda caracterizado por lo que se denomina "frecuencia de muestreo", que representa el número de muestras tomadas por segundo y que, por tanto, se mide en muestras por segundo o hercios (Hz). Calcular el

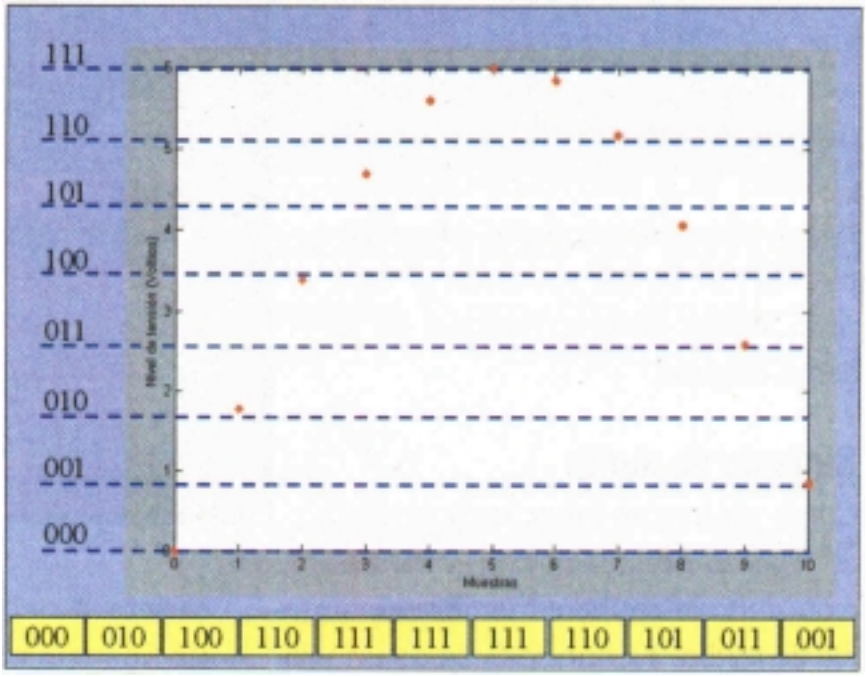


Figura 4. Proceso de cuantización empleando 3 bits. El rango de tensiones disponible (0 a 6 Voltios) se divide en 8 niveles de cuantización (2^3 = 8). A cada nivel de cuantización se le asocia en número de 3 bits. Las muestras de tensión se asocian a los niveles de cuantización. La señal digital resultante se muestra en la parte inferior.

intervalo de tiempo entre una muestra y la siguiente es tan simple como calcular 1 / Fm (donde Fm es la frecuencia de muestreo). Como se puede intuir, la representación digital de la señal será más fiel cuanto mayor sea la frecuencia de muestreo. Tanto es así, que si la frecuencia de muestreo fuera demasiado reducida, se perdería información y no sería posible reconstruir la señal a partir de las muestras digitales. En cualquier caso, es importante recordar que la calidad del sonido capturado será mayor cuanto mayor sea la frecuencia de muestreo.

Por otro lado, el proceso de cuantización queda caracterizado por la resolución del CAD, es decir, el número de bits con que se representa cada muestra. A mayor resolución, el dominio de la amplitud se representará con mayor fidelidad. Si se trabaja con resoluciones demasiado pequeñas, la señal reconstruida a partir de las muestras se encontrará distorsionada por la pérdida de información. También es importante recordar que, a mayor resolución, se requiere de mayor espacio para el almacenamiento de la información de audio.

En general, una buena captura de audio se basa en encontrar un buen compromiso entre frecuencia de muestreo y resolución, lo que se traducirá en calidad de audio y espacio de almacenamiento necesario. Como ejemplo práctico, vamos a suponer que capturamos un tema musical de 5 minutos con la grabadora de sonidos de Windows. Supongamos que seleccionamos una frecuencia de muestreo de 44.100 Hz, una resolución de 16 bits, y sonido estéreo (lo que

implica capturar sonido de dos canales diferentes, simultáneamente), todo ello editando las propiedades del fichero WAV. Ésta es una selección de parámetros oportuna para grabar sonido procedente de un CD-audio. Teniendo en cuerita dichos datos, se almacenarán: 2 canales x 44.100 muestras/s x 16 bits/muestra = 1.411.200 bits por segundo, lo que son 172 kB por segundo, aproximadamente: Si el tema musical dura 5 minutos (300 segundos), se almacenarán 300 s x 172 kB/s = 51:600 kB, y por tanto el fichero WAV resultante ocupará unos 50 MB. Este tamaño de archivo no es excesivo para el almacenamiento en el PC, pero resulta prohibitivo para descargar el fichero a través de Internet. Si desea reducir el tamaño del fichero, podría utilizar 11.025 Hz, 8 bits y sonido mono (un solo canal), lo cual corresponde a calidad de teléfono. En ese caso, el fichero ocuparía alrededor de 3 MB, con lo que se haría viable la descarga vía Internet, pero la calidad del sonido seria inadecuada. En resumen, debe seleccionar los parámetros adecuados al tipo de información a capturar. Para capturar música de un CD, debe elegir la primera opción. Para capturar información hablada con calidad telefónica, debería escoger la segunda opción. En cambio, si desea calidad de radio, debería seleccionar 22.050 Hz y 8 bits.

La conversión digital/analógico

El proceso realizado por el CDA es justamente el inverso al que realiza el CAD, como ya hemos observado. Se parte de muestras en formato binario, y éstas se deben convertir en una señal analógica (continua en el tiempo y la amplitud)

El CDA asocia a cada valor binario un nivel de tensión previamente establecido, y genera muestras de tensión utilizando dichos niveles, aplicando un intervalo de tiempo constante entre muestras. La cuestión a resolver es la siguiente: ¿cómo unir una muestra con la que le sucede? En efecto, dicha unión es necesaria para hacer que la señal vuelva a ser continua en el tiempo. Existen muchas técnicas que hacen esto posible. La más sencilla consiste en mantener el nivel de tensión de una muestra hasta que llegue la muestra siguiente. Otras técnicas más complejas emplean la muestra actual y las muestras anteriores para predecir la siguiente muestra. Después de este proceso, la señal aún presenta cierto grado de distorsión. Por ello, se suele aplicar un proceso de filtrado que suaviza la señal. Si la frecuencia de muestreo y la resolución han sido apropiadas, la señal resultante será una buena reconstrucción de la señal original.

Síntesis de audio

Como ya hemos introducido antes, el proceso de creación de información audible recibe el nombre de síntesis de audio. Ya que se trata de un proceso que -inevitablemente- requiere cómputo, el DSP es el elemento clave en dicho proceso, actuando como "sintetizador". Las tres técnicas más extendidas para la generación sonido son la modulación en frecuencia (FM), las tablas de onda (*wave table*) y el modelado físico (*PhM*).

La síntesis FM permite generar señales complejas -que contienen infinidad de componentes frecuenciales- empleando un procesamiento extremadamente

simple. Por ello la síntesis FM se implementa usualmente en las tarjetas de sonido más económicas. Mediante esta técnica se consigue simular el sonido de multitud de instrumentos, pero cualquier usuario detecta claramente que no se trata del instrumento real, sino más bien de una aproximación.

En el caso de las tablas de onda, se capturan pequeñas secuencias de audio tomadas de instrumentos reales. Las señales capturadas de cada instrumento vienen almacenadas en la tarjeta de sonido. Cuando se emula un instrumento, se toma la muestra oportuna de la memoria y se reproduce con distintas velocidades, obteniendo así las distintas notas musicales. Las tablas de onda constituyen el método de síntesis de mayor calidad, pero implican un coste considerable. Un buen ejemplo de síntesis basada en tablas de onda se encuentra en la conocida tarjeta SoundBlaster AWE.

Finalmente, el modelado físico simula los instrumentos mediante algoritmos de cómputo. Este método aplica modelos de simulación de las propiedades físicas de los instrumentos. En concreto, las entidades que se combinan para generar el sonido se denominan excitadores y resonadores. Los excitadores modelizan la causa que provoca la aparición del sonido. Ejemplos son la pulsación de una tecla, golpear un tambor o desviar una cuerda de su posición de equilibrio. Los resonadores modelizan la respuesta del instrumento ante la excitación aplicada, lo que normalmente se traduce en simular la vibración de los componentes físicos del instrumento. Este método se caracteriza por una elevada carga computacional, y por tanto requiere de técnicas adicionales para poder trabajar en tiempo real. Como ejemplo conocido, se puede citar la tarjeta de sonido SoundBlaster Gold, que contiene 14 instrumentos creados a partir de modelos físicos.

Un parámetro importante a tener en cuenta -y que diferencia a unas tarjetas de otras- es el número de notas (voces) ' que pueden sonar simultáneamente. Cuanto más profesional sea una creación musical, mayor número de voces es necesario. Por ejemplo, la tarjeta SoundBlaster AWE64 admite 64 voces, mientras que la tarjeta SoundBlaster 16 tan sólo admite 20.

Una buena forma de comprobar la calidad de síntesis de una tarjeta consiste en reproducir un fichero MIDI. Nótese que dichos ficheros definen la información musical que debe sonar, pero el sonido real lo pone la tarjeta. Por tanto, un mismo fichero sonará mejor en unas tarjetas que en otras, dependiendo de su calidad.

CAPÍTULO 14

El monitor

En el Capítulo 12, se introdujeron las tarjetas de vídeo como el elemento encargado de convertir la información digital generada por la CPU en señales analógicas representando una imagen. En este capítulo, se aborda el componente que termina la cadena, formando finalmente las imágenes: el monitor.

El monitor es un componente al que, habitualmente, se da menor importancia que a otros elementos del PC (CPU, disco duro, etc.). Ocurre algo similar con el ratón, el teclado o los altavoces: realmente, no es en lo primero que se piensa cuando se imaginan las prestaciones de un PC. Se tiende a pensar en dichos componentes como accesorios, cuando en realidad son elementos importantes. Quizá esto ocurre debido a su localización externa al equipo. A mayor distancia de la CPU, los componentes del PC tienden a comportarse como transductores, elementos de interfaz entre el hombre y la máquina.

Por esta naturaleza de interfaz, componentes como los monitores son fáciles de manejar y entender por el ser humano. También resultan familiares a todo usuario, puesto que muchos de ellos son empleados en otros contextos alejados del PC (por ejemplo, los altavoces también se emplean en equipos de audio).

Centrando la atención en los monitores, recuerdan a los omnipresentes receptores de televisión, que forman parte de nuestra vida cotidiana, lo que proporciona la sensación de que no aportan demasiadas novedades. También con frecuencia, las capacidades gráficas del equipo se atribuyen a las tarjetas de vídeo, olvidando un matiz: si se emplea un monitor de calidad inapropiada, no se aprovecharán realmente las virtudes de la tarjeta. Lo mismo pasaría si se conectasen unos altavoces de baja calidad a un excelente equipo de audio. En resumen, no hay que olvidar que es el monitor quien forma final mente las imágenes que vemos.

Vista la importancia real de los monitores en el mundo del PC, en este capítulo vamos a cubrir el funcionamiento y características de dichos dispositivos, centrándonos en los más comunes en equipos de sobremesa: los de tubo de rayos catódicos (CRT).

Es importante destacar que el interior del monitor constituye un terreno muy peligroso. Se manejan tensiones eléctricas muy elevadas que pueden llegar a causar graves daños al ser humano. Por ello, la apertura del monitor debe realizarla siempre una persona cualificada.

Componentes de un monitor

Aunque su funcionamiento es simple desde el punto de vista del usuario, el interior del monitor encierra un sistema complejo. El componente estrella (y el más costoso) es el tubo de rayos catódicos (TRC, ver Figura 1). Éste contiene varios cañones, cuyo cátodo genera electrones, que son acelerados -a través

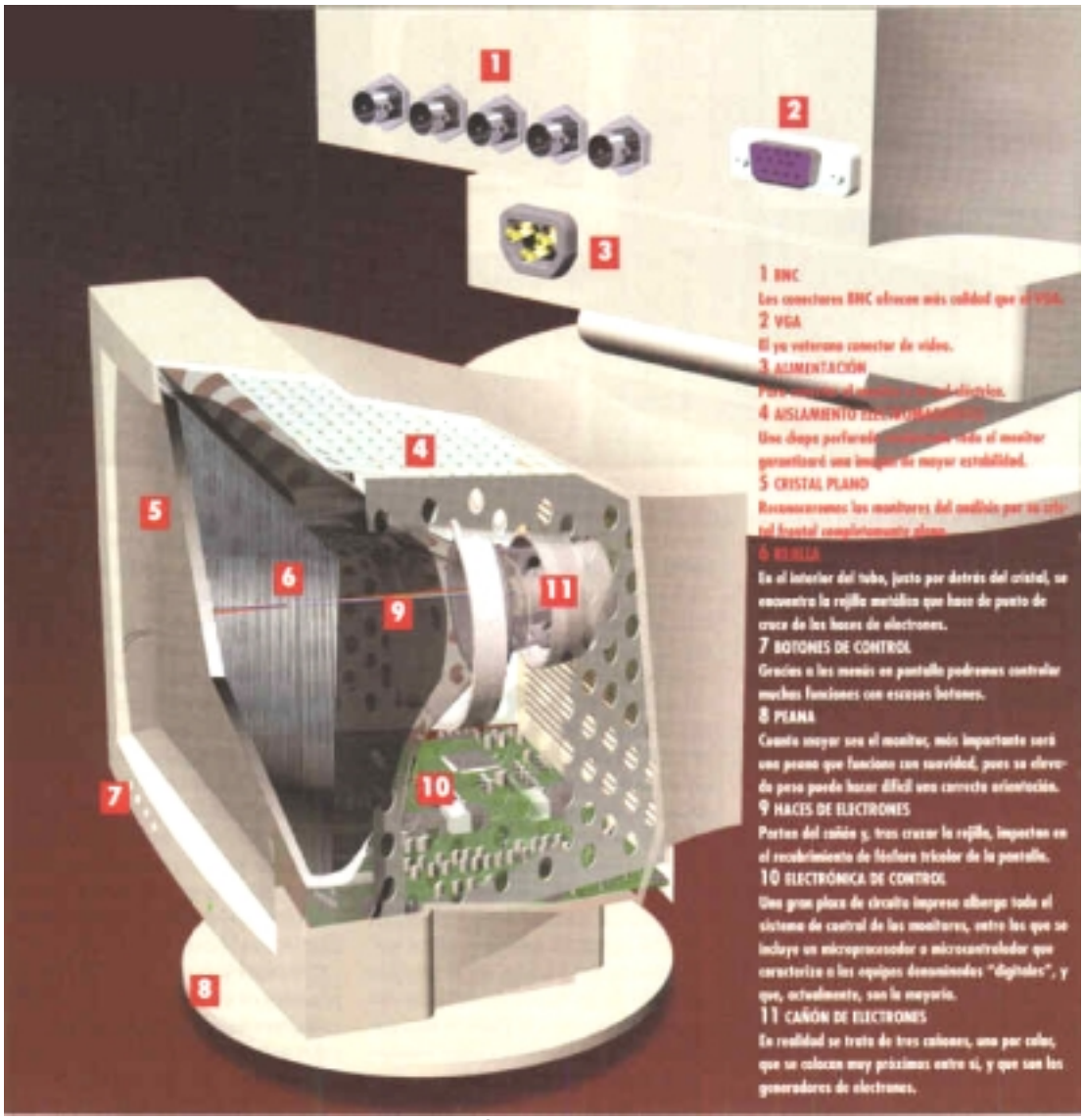


Figura 1. El monitor de tubo de rayos catódicos

del ánodo- hacia un material fosforescente (la pantalla). El cañón barre toda la pantalla, enfocando cada zona sensible y lanzando un haz de electrones con una cierta intensidad.

La pantalla está formada por una serie de zonas sensibles fosforescentes (píxeles), que al ser excitadas por los electrones, emiten radiación visible hacia el usuario. La intensidad de los haces de electrones condiciona la luminosidad de cada píxel, mientras que la composición del fósforo determina su color.

Tras ser excitados, los puntos sensibles de la pantalla son capaces de emitir radiación visible solamente durante un corto periodo de tiempo. Por ello, dichos puntos deben ser excitados de nuevo (léase, refrescados). Esto se consigue realizando el proceso de barrido multitud de veces por segundo. Si la frecuencia de refresco es apropiada, el usuario percibirá una sensación de continuidad de la imagen en el tiempo. En cambio, si dicha frecuencia es demasiado reducida, la pantalla deja de emitir radiación luminosa entre refresco y refresco, haciendo que el usuario perciba un parpadeo en la imagen. Por otra parte, si la frecuencia de refresco es demasiado elevada, el usuario no va a percibir nin-

guna ventaja (no hay que olvidar que el ojo tiene su propia frecuencia de muestreo para capturar imágenes) y, además, se requerirá un elevado ancho de banda entre la tarjeta de vídeo y el monitor para mover tanta información por segundo. Por tanto, la elección de la frecuencia de muestreo está sujeta a un compromiso.

El TRC está gobernado por un circuito controlador. Éste recibe las señales analógicas procedentes de la tarjeta de vídeo y controla al TRC en consecuencia, haciendo que las imágenes se formen sobre la pantalla.

El monitor también dispone de componentes de interfaz con el usuario, que se materializan en forma de controles situados en el exterior del monitor. Estos también se hallan conectados al circuito controlador del monitor, que es quien se encarga de hacer efectivas las órdenes del usuario. Los controles del monitor permiten modificar parámetros como el brillo, el contraste, etc. Respecto al suministro de energía eléctrica, el monitor es el único componente estándar del PC que dispone de su propia fuente de alimentación. Algunos equipos disponen de un zócalo extra, que permite conectar el cable de alimentación del monitor directamente sobre el PC. Esto no significa que el monitor reciba energía de la fuente interna del PC. En realidad, el PC deja pasar su alimentación de corriente alterna -procedente de la red eléctrica- hacia el monitor. La ventaja radica en que, al conectar/desconectar el PC, el monitor se conecta/desconecta automáticamente. Otro aspecto fundamental es la interfaz con el PC, que permite a la tarjeta de vídeo enviar las señales analógicas necesarias para el gobierno del monitor.

Todavía queda por introducir un último componente: la cubierta del monitor. En el caso del monitor, su papel protector es importante, ya que, como se ha dicho antes, oculta un hardware peligroso para el usuario. Además, hay que recordar que los componentes internos del monitor generan una gran cantidad de calor. Por ello la cubierta contiene multitud de ranuras, que aseguran una correcta ventilación. Es importante evitar la obstrucción de dichas ranuras; de lo contrario, el monitor podría calentarse en exceso y acabar averiándose.

Funcionamiento del TRC

Como ya hemos introducido, la misión fundamental del cañón es barrer toda la pantalla, dotando de un color e intensidad luminosa a cada píxel. Este proceso es imprescindible, y debe repetirse varias veces por segundo (como dato práctico, las frecuencias de refresco estándares son 56, 60, 65, 70, 72, 75, 80, 85, 90, 95, 100, 110 y 120 Hz).

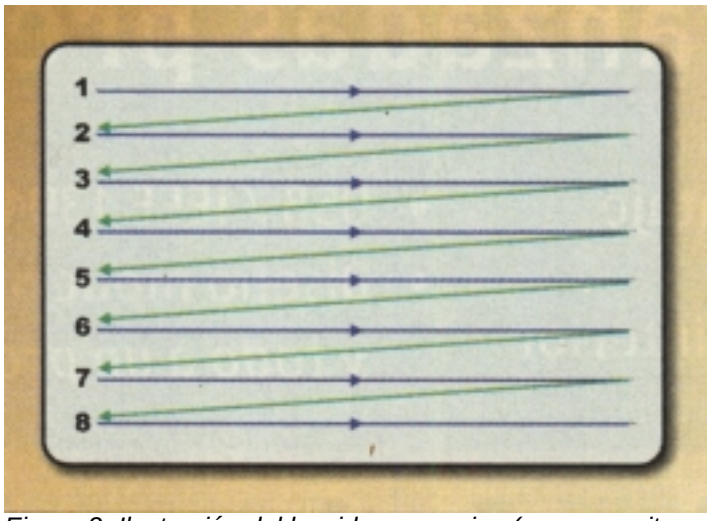


Figura 2. Ilustración del barrido progresivo (en un monitor imaginario de 8 líneas)

En primer lugar, se comienza en el píxel situado en la parte izquierda superior de la pantalla. Entonces, se barren todos los píxeles de la línea superior en sentido horizontal, de izquierda a derecha. A continuación, el haz se desactiva, y el cañón se desplaza hacia el primer píxel de la línea inmediatamente inferior (como si de un “retorno de carro y avance de línea” se tratara). El proceso se repite hasta cubrir toda la pantalla. Finalmente, el haz se vuelve a desactivar, y el TRC vuelve a enfocar al píxel original, listo para “dibujar” una nueva pantalla. Este proceso se denomina “barrido progresivo”, y queda ilustrado en la Figura 2.

Existe otro tipo de barrido, denominado “entrelazado”, que se emplea en el mundo de la televisión, y que también se utilizaba en los primeros monitores, para aprovechar los desarrollos existentes. Mediante esta técnica, en cada refresco sólo se rellena la mitad de las líneas de la pantalla. En un primer barrido, se rellenan las líneas impares (-a). En el barrido siguiente, se rellenan las pares (Figura 3-b), completando un cuadro. El barrido entrelazado tiene una clara motivación: por diversas causas (siempre dentro del mundo de la TV), no es posible ofrecer frecuencias de refresco suficientemente altas. Usando barrido progresivo con frecuencias de refresco insuficientes, se produce una sensación de parpadeo que, a su vez, se convierte en fatiga visual tras varias horas de trabajo. Se podría pensar en aumentar la persistencia de la pantalla, pero esto produciría estelas, especialmente ante movimientos rápidos.

Con el barrido entrelazado se duplica la frecuencia de refresco, utilizando el mismo ancho de banda para transmitir las señales; Aunque cada semi-pantalla (denominada “campo”) contiene la mitad del total de líneas de la pantalla, los píxeles emiten luz el tiempo suficiente para que el ojo crea que todas las líneas de la pantalla han sido dibujadas en un barrido completo. En resumen, desaparece el efecto de parpadeo y no se requiere ancho de banda adicional, lo que

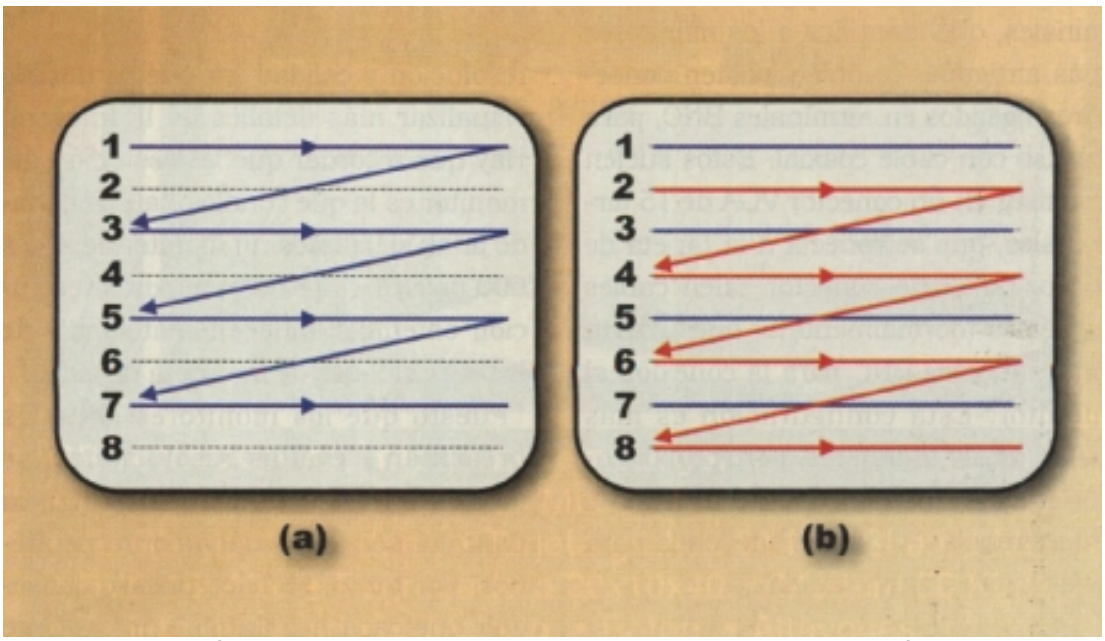


Figura 3. Ilustración del barrido entrelazado (en un monitor imaginario de 8 líneas). Primero se barren las líneas impares (a), y finalmente se completa la imagen con las líneas pares (b).

representa una solución ideal para televisión. Esto funciona muy bien con imágenes en movimiento (típicas en televisión), pero ofrece problemas con imágenes estáticas (el caso del PC). En efecto, si la imagen presenta líneas finas horizontales en una posición fija, se aprecia un efecto de temblor, ya que dichas líneas se refrescan en barridos alternados. Un buen ejemplo son las hojas de cálculo.

Por todo ello, en los PC se emplea barrido progresivo, pero con una frecuencia de refresco bastante superior a la utilizada en televisión. Es posible, por tanto, emplear frecuencias como 72 Hz, que permiten trabajar durante muchas horas con el PC sin fatigar la vista.

El TRC se halla rodeado de bobinas de hilo conductor, denominadas bobinas deflectoras (). Estas bobinas generan campos magnéticos (controlados por la tensión que se les aplica) que actúan sobre los electrones lanzados, modificando su trayectoria. Un bobinado permite modificar la trayectoria de los electrones en sentido horizontal, mientras que otro hace lo mismo verticalmente. De esta forma, mediante la aplicación de dos tensiones eléctricas, se consigue el barrido horizontal, los desplazamientos entre líneas y los saltos al punto inicial de la pantalla.

Utilizando un solo haz, se conseguiría una imagen en escala de grises. En ese caso, que corresponde a los antiguos monitores monocromos, la pantalla se halla recubierta de un material que emite un solo color (normalmente verde, naranja o blanco). Para conseguir el color, se utilizan tres cañones, y los píxeles se forman entrelazando puntos de tres colores distintos, denominados colores primarios. Cada haz se dedica a uno de esos tres colores, y mediante la mezcla aditiva de los mismos (por proximidad), se obtiene cualquier otro color. El sistema empleado en los monitores es el RGB: rojo (R, del inglés red), verde (G, de green) y azul (B, de blue). La intensidad total de los haces determina el brillo de cada píxel, siendo la intensidad relativa entre los tres cañones la que condiciona el color.

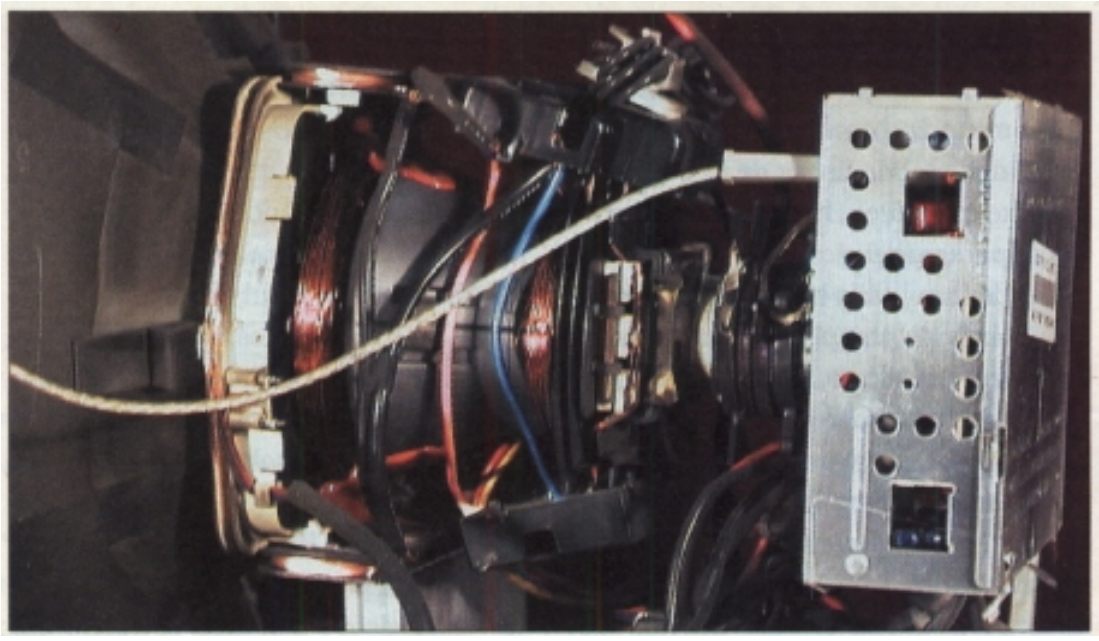


Figura 4. Bobinas deflectoras.

Controles del monitor

En general, existen dos tipos de controles: analógicos y digitales. Los controles analógicos se basan en componentes como las ruedas giratorias (los más habituales en los antiguos televisores). Se caracterizan por su sencillez de uso, y por facilitar un ajuste fino.

En el extremo opuesto, los controles digitales suelen basarse en pulsadores. El control se suele apoyar con imágenes sobre la pantalla (*On Screen Display* o OSD), que suelen contener menús de opciones. De esta forma, el proceso es realmente intuitivo (aunque en un grado que yaría ampliamente de unos monitores a otros). Incluso se suele incorporar memoria, permitiendo almacenar configuraciones para su posterior restauración. Esto último resulta muy útil cuando se cambia de resolución (lo que suele llevar a configurar de nuevo varias características).

El empleo de OSD permite disponer de muchas opciones con un número de controles muy reducido. De hecho, muchos monitores constan tan sólo de un botón y un control giratorio. El control giratorio suele permitir desplazarse entre las opciones de los menús, mientras que el botón permite seleccionar las opciones deseadas o entrar en submenús. La desventaja de esta aproximación radica en que, habitualmente, se requiere cierta práctica hasta familiarizarse con su uso. –

La interfaz con el PC

Como introdujimos en el capítulo 12, la tarjeta de vídeo genera señales analógicas (antiguamente digitales) que controlan el funcionamiento del monitor, haciendo posible visualizar imágenes. En otras palabras, dichas señales son las que gobiernan al TRC. Por ello se hace necesario establecer una interfaz para la interconexión de tarjetas de vídeo y monitores.

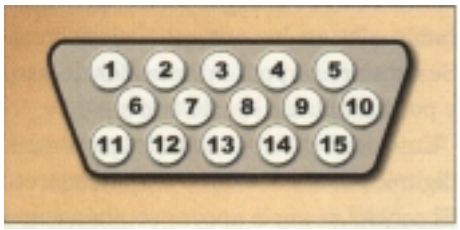


Figura 5. Conector VGA de 15 terminales.

La interfaz empleada en la actualidad se basa en un conector de 15 pines, al que se suele denominar “conector VGA”, puesto que fue el primero en emplearse con dicho estándar. La muestra la disposición de los terminales en dicho conector.

Los terminales 1, 2 y 3 contienen las señales de intensidad correspondientes a los colores primarios (rojo, verde y azul), controlando la intensidad de los haces de electrones correspondientes. Los terminales 6 al 8 son las referencias de tensión (masas) correspondientes a dichas señales.

Los terminales 13 y 14 contienen las señales de sincronía horizontal y vertical, respectivamente. La señal de sincronía horizontal contiene pulsos, que indican el final de una línea de píxeles, es decir, el momento en que el TRC debe retornar hasta el origen de la línea siguiente.

Por otro lado, la señal de sincronía vertical (también pulsante) indica la llegada del haz de electrones al final de la pantalla, o lo que es lo mismo, el momento en que el TRC debe retomar al píxel superior izquierdo. Nótese que todas estas señales se envían de forma separada al monitor. En la televisión, en cambio, dichas señales se empaquetan en una única "señal de vídeo compuesta".

Otro conector conocido es el de 9 terminales, que identifica a los monitores más antiguos. También existen conectores basados en terminales BNC, para su uso con cable coaxial. Estos suelen consistir en un conector VGA de 15 terminales, que se conecta a la tarjeta de vídeo. De dicho conector salen cables coaxiales (normalmente 5) que acaban en terminales BNC para la conexión al monitor. Esta configuración es más cara que las anteriores, pero proporciona el mejor blindaje frente al ruido e interferencias, y es más adecuada para resoluciones muy elevadas.

Máscara de sombra y rejilla de apertura

Tal y como hemos expuesto antes, la pantalla está formada por píxeles, y más aún, por un entrelazado de pequeños puntos, correspondientes a los tres colores primarios (RGB). La distancia entre dos puntos adyacentes correspondientes al mismo color se denomina *dot pitch*, que podríamos traducir como granularidad, y constituye un parámetro muy importante e íntimamente relacionado con la resolución del monitor. A menor *dot pitch*, los puntos son más finos, y por tanto se alcanza una mayor resolución y calidad (ya que es posible visualizar más detalles de la imagen). Hay que recordar que la resolución del monitor es la que condiciona el resultado final: si usamos un monitor de 800 x 600 píxeles, esa será la máxima resolución obtenible, independientemente de las capacidades de la tarjeta de vídeo.

Puesto que los monitores actuales presentan elevadas resoluciones, se deduce que los puntos que forman la pantalla son extremadamente pequeños. Por tanto, se hace preciso conseguir que cada haz ilumine únicamente los puntos dedicados al color correspondiente. Esto se consigue por medio de una placa metálica perforada, que se denomina "máscara de sombra". Las perforaciones se ajustan perfectamente a los puntos existentes en la pantalla. De esta forma, colocando cuidadosamente la máscara de sombra en la posición adecuada, los haces se dirigirán siempre hacia los puntos apropiados.

Algunos monitores consiguen lo mismo empleando lo que se denomina "rejilla de apertura". Este tipo de máscara sustituye las perforaciones por ranuras. Por ello, las áreas iluminadas en la pantalla aparecen oblongas o en forma de tiras. La principal ventaja de esta técnica radica en que permite que un mayor porcentaje de la radiación llegue a la pantalla (los cañones se hallan situados en un mismo plano, en línea con las ranuras), consiguiendo colores de gran pureza. La principal desventaja es una reducida estabilidad física respecto a la máscara de sombra: las ranuras tienden a vibrar. Para eliminar este efecto, se emplean cables estabilizadores muy finos, dispuestos horizontalmente. Esta solución tiene un efecto secundario: la aparición de suaves líneas horizontales justo en la posición de los cables. Dichas líneas no se suelen apreciar a simple vista, a menos que se busquen intencionadamente (con la pantalla totalmente

coloreada en blanco se detectan con mayor facilidad). Un TRC que emplea esta técnica es el popular Trinitron de Sony, empleado en muchos monitores.

Aspectos de seguridad, protección, energía y radiación

Debido a los principios de funcionamiento del TRC, la energía eléctrica con sumida por el monitor es tanta o más que la consumida por el resto del PC. Por ello han surgido estándares que afectan al consumo de energía, exclusivamente dedicados al monitor. La agencia de protección medio-ambiental americana (EPA) lanzó un programa llamado *EnergyStar*, que permite certificar los PC y monitores que siguen una serie de normas que aseguran un apropiado consumo de energía.

Otra iniciativa, presente en la mayor parte de los monitores actuales, es el protocolo DPMS (*Display Power Management System*), un sistema de administrar la energía consumida por el monitor. DPMS permite que ciertos subsistemas del monitor se desactiven tras un cierto periodo de inactividad, consiguiendo reducir el consumo considerablemente. El sistema operativo debe soportar esta característica. Cuando se detecta inactividad durante cierto periodo de tiempo, el sistema operativo envía una señal al monitor, pasando a un modo de bajo consumo. En muchos monitores, existe incluso un modo en el que el monitor se desconecta completamente, ahorrando mucha energía. Cuando se detecta actividad de nuevo, el sistema operativo envía una señal al monitor para retornar al estado normal de trabajo. El único inconveniente de los sistemas DPMS reside en una posible configuración inapropiada: si se conmuta el estado con excesiva frecuencia, los componentes internos resultan afectados, reduciendo la vida útil del monitor (por ejemplo, esto podría ocurrir si se indica pasar a bajo consumo tras sólo un minuto de inactividad).

Las zonas sensibles de la pantalla, tras ser iluminadas de forma continuada, pueden sufrir daño. Esto es particularmente cierto cuando una imagen permanece estática en la pantalla durante un tiempo prolongado: los mismos puntos de la pantalla están siendo activados continuamente. Si se llega al punto de deteriorar la pantalla, quedarán secuelas de la imagen que ha causado el problema, incluso cuando el monitor está apagado. Como solución a este problema, nacieron los salvapantallas que básicamente, son programas que muestran imágenes en movimiento (o ninguna imagen en absoluto) tras cierto periodo de inactividad, evitando a toda costa cualquier imagen estática que pueda producir daños. En realidad, el deterioro de la pantalla era común en los antiguos monitores monocromos, pero apenas ocurre en los monitores modernos. Por tanto, los salvapantallas no son realmente necesarios. De hecho, se han convertido en elementos de entretenimiento, e incluso algunos llegan a mostrar imágenes totalmente estáticas.

De nuevo, debido a la forma de trabajo del monitor, produce emisiones electromagnéticas que -teóricamente- pueden causar daño a los usuarios que permanecen frente al PC un gran número de horas. En consecuencia, también han surgido estándares para la fabricación de monitores con bajos niveles de emisión electromagnética. La pregunta clave es: ¿causan daño, realmente, dichas emisiones? No existe un verdadero acuerdo entre las industrias de la salud y la

informática, por lo que la cuestión queda abierta. En cualquier caso, el mejor consejo es la prudencia: evitar trabajar excesivamente cerca del monitor.

En cuanto a los aspectos de seguridad, no hay que olvidar que el monitor trabaja con tensiones eléctricas muy elevadas, que pueden causar graves daños al ser humano, e incluso la muerte. Este peligro permanece incluso después de haber desconectado el monitor, ya que se emplean multiplicadores de tensión basados en condensadores (léase, almacenes de tensión). Por todo ello, es muy importante no abrir la cubierta del monitor para trabajar sobre su interior. Cualquier trabajo sobre los componentes internos debe ser realizado por personal técnico cualificado.

CAPÍTULO 15

LOS PUERTOS DEL PC

A través de dos capítulos anteriores se presentaron las interfaces IDE, SCSI, ISA, PCI y AGP corno medios para la conexión de periféricos al PC.. Este capítulo complementa la trilogía, abordando los principales canales estándares para la conexión de periféricos externos al PC: los puertos serie, paralelo, USB y *FireWire*.

Las ranuras de expansión del PC no son la única opción a la hora de conectar periféricos al PC. De hecho, los periféricos externos suelen precisar la instalación de una tarjeta adaptadora (normalmente PCI), a la cual se conecta el dispositivo en sí (disminuyendo, por tanto, el número de ranuras disponibles y limitando la capacidad de expansión). Afortunadamente, cualquier PC actual está equipado con diversos puertos para la comunicación directa con dispositivos externos (ratón, impresora, *webcam*, asistente personal, etc.), sin necesidad de instalar tarjetas adaptadoras. Cada tipo de puerto está dotado de unas características y un funcionamiento diferentes, que condicionan el tipo de dispositivos que se pueden conectar. Por ejemplo, la conexión de un módem sugiere el empleo de un puerto serie, mientras que una impresora invita a la utilización del puerto paralelo. Este capítulo enfoca el funcionamiento de los puertos externos más conocidos y de mayor presencia en el mundo del PC, en la actualidad. En primer lugar, se abordarán los puertos serie y paralelo, dos “clásicos” que todo PC incorpora prácticamente “por definición”, y que se emplean con frecuencia para la conexión de ratones, módems e impresoras. A continuación, hablaremos de dos canales de expansión muy actuales y que gozan de una excelente aceptación: los buses USB y *FireWire*.

Puertos serie

Los puertos serie –también conocidos como puertos de comunicaciones (COM) están considerados como una interfaz externa fundamental. De hecho, dichos puertos han acompañado al PC desde hace más de veinte años. En general, todo PC incluye dos puertos serie RS - 232, denominados COM1 y COM2. En la actualidad, los fabricantes tienden a emplear medios de conexión más modernos, como el bus USB. Pero, sin embargo, todavía existe multitud

Figura 1. Conector DB9 –en sus variantes macho y hembra205- y sus correspondientes señales. Nótese que la numeración de pines es diferente en cada variante

de dispositivos diseñados para trabajar a través del puerto serie, incluyendo módems, equipos de medida, receptores GPS, plataformas de sincronización para PDA, etc.

En general, una característica básica del puerto serie hace referencia a la velocidad de transferencia de datos que es capaz de ofrecer: muy reducida. La mayoría de puertos serie son capaces de ofrecer relaciones de transferencia de hasta 115 kbps. En consecuencia, el puerto serie resulta una elección acertada para la comunicación a velocidades no muy exigentes. Por ejemplo, el funcionamiento de un ratón exige enviar información al PC a una velocidad nada llamativa en comparación con muchos otros periféricos. Por tanto, emplear un puerto serie es una solución más que suficiente, y de hecho es la solución típica (emplear un canal más rápido implicaría desaprovechar sus posibilidades).

Señales empleadas por el puerto serie

Los conectores DB9 y DB25 (Figuras 1 y 2), a pesar de presentar un diferente número de terminales transportan los mismos tipos de señales. El uso principal para el que fue diseñado el puerto serie consistía en la conexión de un módem, hecho que se refleja claramente en la disposición de los terminales. En primer lugar, se encuentra el cable dedicado al envío de datos en serie hacia el módem (Transmit Data, TxD), así como el correspondiente a la recepción de datos procedentes del módem (Receive Data, RxD). Para inicializar las comunicaciones, el módem emplea la señal Data Set Ready (DSR) para comunicar que éste se encuentra preparado para iniciar el proceso de intercambio de datos. De forma análoga, la UART utiliza la señal Data Terminal Ready (DTR) para indicar que el PC se encuentra listo.

Una vez iniciada la comunicación serie, la UART envía la señal Request to Send (RTS) al módem para consultar si éste está preparado para recibir información. El módem utiliza la señal Clear to Send (CTS) para contestar, indicando que la UART puede enviar datos. Generalmente, los módems actuales trabajan a 56 kbps, mientras que la conexión entre PC y módem suele ser mucho más rápida (típicamente 115 kbps). Aun teniendo en cuenta que el módem dispone de una memoria para almacenar los datos procedentes del PC, dicha memoria se llena muy rápidamente, mientras que el módem procesa los datos (es decir, vacía la memoria) con mayor lentitud. Es ahí donde se centra la utilidad principal de las señales RTS y CTS: el módem puede detener la recepción de datos, y reanudarla cuando es preciso, de forma que la memoria del módem no se desborde.

La señal Ring !ndicator (Rl) se emplea para detectar la recepción de una llamada. Por otro lado, la señal CarrierDetect (CD) indica si el módem se halla conectado a una línea telefónica.

Todas las señales arriba expuestas presentan una naturaleza digital, y por tanto sólo pueden presentar dos estados lógicos (“1” o “0”). Todas las señales se hallan referidas a una misma referencia de tensión (masa), accesible mediante el terminal Signa! Ground (GND).

Básicamente, el puerto serie define un conector y un protocolo para el Intercambio de información. Tal y como su nombre indica, la información se transmite y recibe en serie. En otras palabras, toda la información a Intercambiar circula por un único cable, moviendo un bit en cada ciclo de transferencia. Por tanto, para enviar una palabra digital de 8 bits, se enviará un bit tras otro, cubriendo un total de 8 ciclos de transferencia. La ventaja fundamental radica en que sólo es necesario un cable para el intercambio de información, lo que reduce costes. La desventaja principal ya ha sido introducida: la velocidad de transferencia es reducida. De hecho, si se emplearan 8 cables en lugar de uno, la transferencia de un byte requeriría tan sólo de un ciclo de reloj (léase, se trabajaría 8 veces más rápido). Es importante destacar que los puertos serie son bidireccionales, es decir, permiten enviar y recibir información simultáneamente. Por ello, realmente existen dos cables dedicados al intercambio de información uno de ellos se emplea para enviar datos y otro para recibirlos. El funcionamiento del puerto serie se implementa, al completo, mediante un chip llamado

UART (Universal Asynchronous Receiver/Transmitter). Este chip toma palabras digitales procedentes del bus del sistema, las convierte a formato serie, y las envía al dispositivo de destino aplicando el protocolo pertinente. A su vez, la UART recibe los datos serie del dispositivo externo, y los entrega al sistema en forma de palabras digitales. Por tanto, la CPU no se debe preocupar de los detalles del protocolo de envío / recepción, quedando libre de dicha carga. La CPU tan sólo entrega la información a enviar a la UART y ésta se encarga de hacer efectivo el envío serie. Para leer datos, se acude a la UART en el momento deseado -que no tiene por qué ser el momento en que los datos están siendo enviados por el periférico externo-. Este modo de trabajo exige la existencia de dos elementos de memoria (buffers): uno se emplea para escribir la información que la UART debe enviar, mientras que al otro se acude para obtener los datos recibidos. Esto permite, por tanto, escribir los datos a enviar mientras se recibe información, y viceversa. La capacidad de dichos elementos de memoria suele oscilar entre 16 y 64 kB.

Figura 2. Conector DB-25 –en sus variantes macho y hembra- y sus correspondientes señales. Nótese que la numeración de los pines es diferente en cada variante. Las señales disponibles son las mismas que en el caso del conector DB-9

Los conectores correspondientes al puerto serie se presentan en versiones de 9 y 25 terminales (ver Figuras 1 y 2), cuyas denominaciones estándares son DB-9 y DB-25, respectivamente.

El puerto paralelo

Las impresoras recuerdan inmediatamente la imagen mental del puerto paralelo, ya que es ésta la interfaz mayormente empleada para la conexión de dicho periférico. Durante el diseño de los primeros PC, IBM introdujo dicho puerto, con el objetivo de conectar una impresora. Además de las impresoras, el puerto paralelo ha sido un medio eficaz para la conexión de muchos otros periféricos, como escáneres, algunas grabadoras de CD, discos duros externos, discos ZIP, etc.

El funcionamiento del puerto paralelo se basa en el envío de un byte completo en cada transferencia, siendo necesarios, por tanto, 8 cables dedicados al intercambio de información. El puerto serie necesita 8 operaciones de transferencia para enviar un byte, lo cual sugiere que el puerto paralelo puede trabajar

a una velocidad notablemente superior. Como dato práctico, el puerto paralelo estándar alcanza velocidades entre 50 y 100 kB por segundo.

El conector vuelve a ser el DB-25 (ver Figura 2). Las señales disponibles se muestran en la Figura 3. Otra variante muy conocida es el conector Centronics de 36 terminales (Figura 4) que, a pesar de la diferencia en el número de terminales, presenta las mismas señales que el conector DB-25.

Terminal	Señal
1	Strobe
2	Data0
3	Data1
4	Data2
5	Data3
6	Data4
7	Data5
8	Data6
9	Data7
10	Acknowledge
11	Busy
12	Paper End
13	Select
14	Auto Feed
15	Error
16	Init
17	Select In
18 - 25	GND

Figura 3. Señales empleadas por el puerto paralelo, y sus correspondientes terminales dentro del conector DB-25

El puerto paralelo original era unidireccional, y por tanto las señales viajaban desde el PC hacia la impresora, nunca en el sentido opuesto. Tras el lanzamiento del PS/2, IBM ofreció una nueva versión del puerto paralelo (denominado Standard Parallel Port o SPP). Este nuevo diseño era bidireccional, y consiguió reemplazar al puerto paralelo original. En principio, los terminales 2-9 se usan para el envío de datos, lo que implica que en cada transferencia los datos viajan del PC al dispositivo externo o viceversa, pero no es posible la transferencia simultánea en ambos sentidos. En otras palabras, la comunicación es *half duplex.* Por fortuna, los terminales 18 al 25, ori-

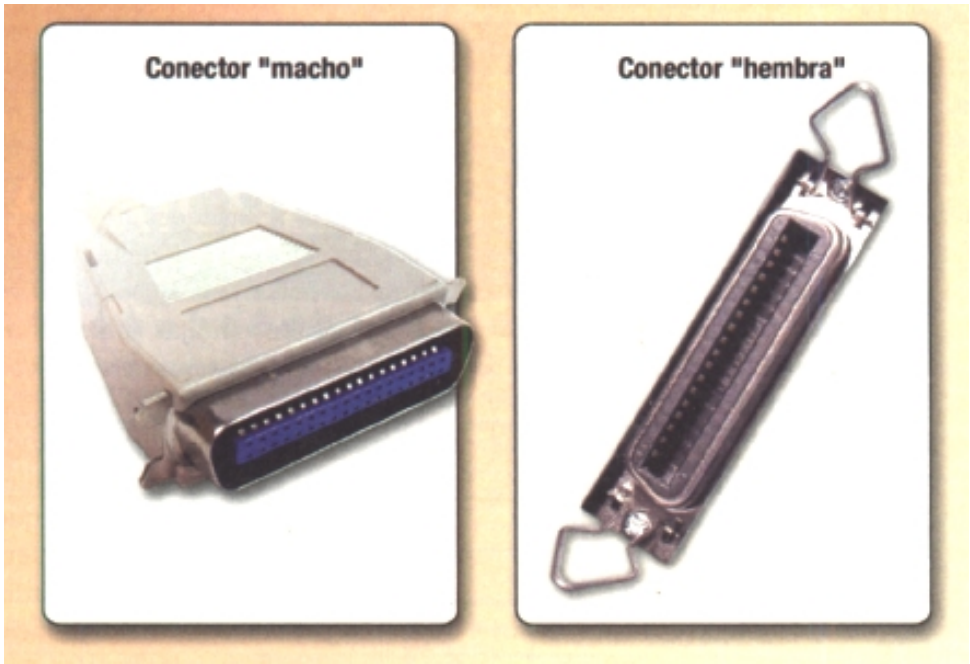


Figura 4. Conector Centrónicx de 36 terminales (variantes macho y hembra)

ginalmente empleados como masas, pueden usarse también como terminales de datos, permitiendo la comunicación bidireccional simultánea (*full-duplex*).

Con el paso del tiempo, nuevas variantes del puerto paralelo han ido apareciendo, como el puerto EPP (Enhanced Parallel Port), que permite enviar entre 500 kb y 2 Mb de datos cada segundo.

Señales empleadas por el puedo paralelo

Al igual que las señales del puerto serie están orientadas al uso de un módem, las señales del puerto paralelo se encuentran particularmente en línea con el control de una impresora A continuación se describe el propósito de las señales mostradas en la Figura 3.

Strobe. Esta señal produce un cambio de tensión cada vez que el PC envía un byte de datos De esta forma la impresora detecta que ya se han establecido los estados lógicos deseados en las 8 líneas de datos y por tanto se puede leer la información (un byte).

Data0 Data7. Contienen los datos enviados a la impresora (8 bits cada vez).

Acknowledge. La impresora indica que ha recibido el byte enviado y esta lista para recibir un nuevo byte.

Busy. Al igual que ocurría en el puerto serie (señales CTS y RTS) la impresora utiliza esta señal para indicar que aun no esta preparada para recibir mas datos.

Paper end. La impresora indica que se ha quedado sin papel.

Select. La impresora indica que se encuentra en línea.

Auto Feed. Cuando se recibe un retorno de carro existen dos formas de interpretarlo Algunas impresoras simplemente retornan al comienzo de la línea. Sin embargo, otras añaden -además- un avance de línea La señal auto feed permite conmutar entre dichos modos de funcionamiento.

Error. La impresora emplea esta señal para notificar cualquier tipo de error detectado.

Init. Ante un cambio de estado lógico en esta señal la impresora se reinicializa.

Select In. Esta señal permite poner a la impresora fuera de servicio.

GND. Terminales de masa (referencia de tensión para todas las demás señales).

El bus USB

A pesar de que han persistido desde los inicios del PC, y de su conveniencia para multitud de aplicaciones, los puertos serie y paralelo presentan claras limitaciones en cuanto a capacidad de expansión y rendimiento se refiere. A mediados de los 90, un consorcio formado por multitud de empresas -incluyendo Compaq, IBM, Microsoft, NEC, etc. - desarrolló una nueva interfaz estándar para la conexión de dispositivos externos al PC. Dicha interfaz, bautizada como USB (Universal Serial Bus, Bus Serie Universal), tenía como objetivo conectar periféricos relativamente lentos (ratones, impresoras, cámaras digitales, unidades ZIP, etc.) de una forma realmente sencilla, rápida y basada en comunicaciones serie. El éxito de la interfaz USB ha sido tal que, actualmente, resulta difícil imaginar un PC que no disponga de al menos un puerto USB. Como prueba adicional de dicho éxito, cabe destacar que actualmente la gran mayoría de periféricos existentes en el mundo del PC están disponibles también en versión USB (por ejemplo, en www.usbgear.com están disponibles una infinidad de dispositivos USB).

Una importante característica de los puertos USB es la sencillez con la que se instala un dispositivo. Tan sólo hay que conectar un extremo del cable USB al periférico, y el extremo opuesto se inserta directamente sobre un conector USB, situado en la parte exterior del PC. No es necesario instalar ninguna tarjeta adaptadora ISA o PCI, lo que ahorra multitud de esfuerzo y problemas.

El bus USB ha sido concebido teniendo en mente la filosofía Plug & Play. Por tanto, tras conectar el dispositivo al bus USB, el sistema operativo se encarga

del resto: si el software controlador se encuentra instalado, lo emplea directa mente, y en caso contrario lo solicita.

El bus USB admite la conexión de hasta 127 dispositivos, algo impensable usando puertos serie o paralelo. Cada dispositivo puede trabajar con un ancho de banda de hasta 6 Mbps, velocidad más que suficiente para la mayoría de periféricos. El ancho de banda total soportado por el bus es de 12 Mbps, a repartir entre todos los dispositivos conectados (incluyendo al propio PC, que actúa como dispositivo anfitrión). De ahí se deduce que si se trabaja a 6 Mbps, tan sólo se podrá conectar un dispositivo al bus.

Otra importante característica es que los dispositivos se pueden conectar y desconectar sin necesidad de apagar el PC (hot-swapping).

El estándar USB define dos tipos de conectores, denominados “A” y “B” (Figura 5). En cuanto a terminales se refiere, las características de ambos son totalmente análogas. La diferencia radica en que los conectares tipo “A” llevan la información desde los dispositivos hacia la computadora, y los conectares tipo “B” llevan la información en sentido opuesto. Esta diferenciación evita toda confusión al efectuar conexiones: los conectores tipo “A” irán a parar a receptáculos tipo “A”, y lo equivalente para los conectores tipo “B”, sin miedo alguno a realizar conexiones incorrectas.

El bus USB permite el flujo de información en ambos sentidos (del PC a los dispositivos o de los dispositivos hacia el PC), lo que abre un gran abanico de posibilidades de expansión. Entre otras, es posible conectar altavoces compatibles USB para escuchar audio procedente del PC (sin necesidad de emplear una tarjeta de sonido) o recibir información del mundo físico que rodea al PC mediante una tarjeta de adquisición de datos USB.

Expandiendo el bus USB

La mayoría de los PC actuales incorporan “de serie” una o dos ranuras USB, lo que permite la conexión de uno o dos dispositivos. Considerando la elevada oferta de dispositivos USB existente en el mercado, es fácil que en muy poco tiempo ya se hayan agotado las ranuras disponibles. Entonces, ¿cómo se puede ampliar el número de ranuras, siendo posible llegar hasta 127?. La respuesta la proporciona un dispositivo económico, que forma parte del estándar USB y se denomina comúnmente *hub* USB (Figura 5). Este contiene un terminal que se conecta al PC (o a otro *hub*), y varios terminales (normalmente 4, aun que pueden ser más) que permiten conectar dispositivos USB adicionales. Cualquiera de estos últimos dispositivos puede ser otro *hub*, creando una jerarquía multicapa que incrementa el número de conectores disponibles.

Los hub USB pueden ser alimentados o no alimentados. Los dispositivos conectados a estos últimos toman la alimentación del propio bus USB, lo que simplifica el diseño. Estos dispositivos se caracterizan por consumir poca energía (por ejemplo, ratones o cámaras digitales). Los dispositivos con mayores requerimientos de energía (impresoras, escáneres, etc.) incorporan su propia fuente de alimentación, ya que la energía que puede proporcionar el bus USB no es suficiente. Si se conecta un número elevado de dispositivos alimentados directamente del bus, es probable que sea necesario proporcionar energía ex-

terna para lograr una correcta alimentación. En ese caso, será preciso emplear los hub alimentados, que vienen acompañados de una fuente de alimentación externa.

Funcionamiento del bus USB

Un buen punto de partida para abordar este tema es el cableado del bus. Cada cable USB contiene, a su vez, 4 cables en su interior. Dos de ellos están dedicados a la alimentación (5 voltios) y la referencia de tensión (masa). La corriente máxima que el bus puede proporcionar es de 500 mA a 5 voltios de tensión. Los dos cables restantes forman un par trenzado, que transporta la información intercambiada entre dispositivos, en formato serie.

Tras su encendido, el dispositivo anfitrión -el PC- se comunica con todos los dispositivos conectados al bus USB, asignando una dirección única a cada uno de ellos (este proceso recibe el nombre de "enumeración"). Además, el PC consulta qué modo de transferencia desea emplear cada dispositivo: por interrupciones, por bloques o en modo isócrono.

Figura 5. Elementos fundamentales del estandar USB: conectores tipo "A" y "B", y hubs USB.

La transferencia por interrupciones la emplean los dispositivos más lentos, que envían información con poca frecuencia (por ejemplo teclados, ratones, etc.).

La transferencia por bloques se utiliza con dispositivos que mueven grandes paquetes de información en cada transferencia. Un ejemplo son las impresoras.

Finalmente, la transferencia isócrona se emplea cuando se requiere un flujo de datos constante y en tiempo real, sin aplicar detección ni corrección de errores. Un ejemplo es el envío de sonido a altavoces USB. Como se puede intuir, el modo isócrono consume un ancho de banda significativo. Por ello el PC impide este tipo de transferencia cuando el ancho de banda consumido supera el 90% del ancho de banda disponible.

Para la temporización, el bus USB divide el ancho de banda en porciones, controladas por el PC. Cada porción mueve 1.500 bytes, y se inicia cada milisegundo. Ante todo, el PC asigna ancho de banda a los dispositivos que emplean transferencias isócronas y por interrupciones, garantizando el ancho de banda necesario. Las transferencias por bloques emplean el espacio restante, quedando en última prioridad.

La versión 2.0 de USB

Todo lo comentado en los anteriores apartados corresponde a la versión 1.1 del bus USB. La versión actual del estándar USB es la 2.0. En primer lugar, esta nueva versión es totalmente compatible con la versión 1.1. Por tanto, desde el punto de vista del usuario no hay cambios: los dispositivos para la versión 1.1 seguirán funcionando sin problemas. Se puede, por tanto, combinar periféricos para ambas versiones, pero hay que tener en cuenta que, en caso de emplear dispositivos USB 2.0, será necesario introducir *hubs* USB 2.0, además de los USB 1.1.

La ventaja para el usuario aparece al utilizar dispositivos diseñados para la versión 2.0: el ancho de banda aumenta hasta un factor 40 (lo que implica alcanzar 480 Mbps). Esto hace posible ampliar el abanico de periféricos USB disponibles, siendo posible conectar dispositivos con elevados requerimientos de ancho de banda, como discos duros, grabadoras de CD, lectores DVD, etc. De hecho, ahora es posible trabajar con periféricos de alto rendimiento sin necesidad de emplear interfaces como SCSI, aprovechando así todas las ventajas de USB y reduciendo costes.

Otra ventaja interesante se centra en la concepción de los dispositivos USB 2.0: no requieren esfuerzo de diseño adicional respecto a la versión 1.1.

Todas estas características anticipan -cada vez más- que el PC del futuro tan sólo necesitará conectores externos USB. A corto plazo, la tendencia apunta a que USB 2.0 reemplazará completamente a USB 1.1 en todo PC.

La intertaz FireWire

El término *FireWire* resulta familiar, sobre todo, para los usuarios de PC interesados en el campo del vídeo digital. Pero, más allá de este campo concreto, se trata de un bus serie similar al USB, que admite la conexión de una gran variedad de dispositivos.

El bus *FireWire* fue introducido por Apple (con antelación a USB), y más tarde fue estandarizado bajo la especificación IEEE 1394, referido como un bus serie de altas prestaciones. *FireWire* alcanza velocidades de transferencia de 400 Mbps y permite la conexión de hasta 63 dispositivos.

La mayoría de ventajas comentadas para USB están presentes en *FireWire* (Plug & Play, conexión/desconexión sin apagar el PC, alimentación incluida en el bus, etc.). Una primera diferencia se encuentra en el cable, que empaqueta un total de 6 cables internos (2 para alimentación, y dos pares trenzados para datos). Otra diferencia fundamental hace referencia a la topología del bus: en lugar de emplear *hubs*, se emplea una configuración “en cadena”. En otras palabras, los dispositivos se unen uno a otro formando una cadena, en la cual es posible insertar más de un PC (haciendo posible que varias computadoras accedan a los dispositivos conectados).

En términos de velocidad de transferencia, *FireWire* supera con creces a USB 1.1, pero es muy similar a USB 2.0. *FireWire* esta orientado a dispositivos con

elevados requerimientos de ancho de banda. En cambio, no resultaría rentable fabricar dispositivos lentos para este bus, algo que lo pone en desventaja respecto a USB 2.0 (que admite ambos tipos con un reducido coste). En términos de coste, hay que señalar que la implementación de *FireWire* resulta más cara que en el caso de USB.

Finalmente, es muy importante notar que *FireWire* trabaja con una filosofía *peer-to-peer*, lo que significa que -en oposición a USB- no precisa de la presencia de un dispositivo anfitrión (el PC). Por ejemplo, es perfectamente posible interconectar dos-cámaras mediante *FireWire* sin necesidad de un PC.

Aunque el éxito de USB sobre *FireWire* ha quedado claramente patente, en el terreno del vídeo digital la situación se invierte: la mayoría de cámaras digitales presentes en el mercado incorporan una ranura *FireWire*. Una configuración habitual para los expertos en dicho campo se compone de una cámara, un disco duro y un PC, conectados a través de *FireWire*. La información persiste en toda la cadena en formato digital y viaja a gran velocidad, por lo que no se pierde calidad y el rendimiento es asombroso.

CAPÍTULO 16

LOS MÓDEMS

Este capítulo se sumerge en el funcionamiento del componente que hace posible la conexión a Internet en la mayoría de los hogares de usuarios de PC. Se trata del módem, un componente que ejerce el papel de interfaz entre el PC y las redes telefónicas.

Cada día que pasa, el número de usuarios de PC que dispone de conexión a Internet aumenta vertiginosamente. Hoy resulta casi impensable instalar un PC en un hogar y no dotarlo de acceso a las grandes posibilidades de la red de redes. Pero, ¿por qué vía conectamos nuestro PC a los sistemas que dan acceso a Internet, permitiéndole interactuar con ordenadores situados en cualquier parte del mundo? La mayoría de hogares suelen están conectados a una o varias redes que pueden hacer posible dicha conexión (red eléctrica, televisión por cable, etc.). Entre ellas, la red telefónica ha constituido el medio principal de conexión a Internet. Llegados a este punto, es necesario adquirir un periférico que permita adaptar la información que genera o recibe nuestro PC (de naturaleza digital) a las características de la red telefónica (que conduce señales analógicas en una banda de frecuencias audibles). Dicho dispositivo recibe el nombre de módem, y su importancia actual es tal que viene instalado de fábrica en muchos PC. Hay que tener en cuenta que la red telefónica limita notable mente la velocidad de transferencia de información con respecto a las grandes exigencias actuales (juegos en línea, videoconferencias, etc.). También hay que recordar que actualmente existen vías de conexión más rápidas al alcance de los usuarios (redes de fibra óptica, ADSL, RDSI, cable, GSM, GPRS, etc.). Aun así, no hay que olvidar que el número de usuarios conectados a través de la red telefónica básica sigue siendo considerable. Además, el dispositivo de adaptación a todas estas redes alternativas se sigue denominando módem. Por todo ello, el módem constituye una pieza imprescindible para la conexión a Internet (o en términos más generales, para la interconexión de dos computa doras a través de la red telefónica). Una vez sentada la misión clave del módem -y su íntima relación con Internet- este artículo presenta sus características principales y su funcionamiento.

Introducción al módem

La misión fundamental de un módem es hacer posible el intercambio de datos en tre dos ordenadores a través de la red telefónica. Los módems fueron desarrolla dos y usados por la defensa aérea norteamericana durante los años 50. Sin embargo, los primeros módems comerciales se lanzaron en los años 60. El primero recibió el nombre de Bell 103, desarrollado por AT&T, y funcionaba a 300 bps. El objetivo era interconectar terminales (dispositivos con poco más que teclado, pantalla, y un hardware mínimo sin capacidad de cómputo alguna, lo que coloquialmente se denomina «terminal tonto") a computadoras (grandes dispositivos, que a menudo ocupaban habitaciones enteras y que proporcionaban la potencia de cálculo). Las computadoras

ban la potencia de cálculo). Las computadoras podían estar situadas en cualquier lugar, con la única condición de disponer de una línea telefónica operativa. Era la época de los sistemas de tiempo compartido. Por ejemplo, una compañía podía contratar tiempo de acceso a la computadora, y emplear un módem de 300 bps para conectar sus terminales. Ya que la información intercambiada era básicamente texto, dicha velocidad resultaba más que suficiente.

Con el nacimiento de los ordenadores personales a finales de los años 70, aparecieron también los BBS (*Bulletin Board Systems*). Cualquier usuario podía crear un BBS en casa, con uno o dos módems, y el software apropiado. Los usuarios se conectaban al BBS mediante un módem, y ejecutaban emuladores de terminal, que convertían a sus ordenadores en «terminales tontos" (el procesamiento se efectuaba realmente en el ordenador que implementaba la BBS). En aquel contexto, los módems trabajaban a 300 bps. Esto no suponía un cuello de botella, ya que se intercambiaba principalmente texto y pequeños programas. Con el paso del tiempo, las transferencias típicas incorporaron imágenes y programas de mayor volumen, y los módems fueron adaptándose a este hecho, incrementando su velocidad de transferencia. Hoy, el módem original ha evolucionado hasta los 56 kbps (velocidad difícil de superar por la propia naturaleza de la red telefónica, lo que ha motivado el éxito de nuevas tecnologías como ADSL).

Bits por segundo y baudios

En la literatura sobre módems, los términos "bits por segundo" (bps) y "baudios" aparecen con gran frecuencia, como medida de la velocidad de transferencia que el módem puede ofrecer. Muchos usuarios tienden a interpretar que ambos términos significan lo mismo, cuando esto no es realmente así.

La relación de transferencia en bps queda totalmente autodefinida: es el número de bits que se pueden enviar cada segundo. En cambio, los baudios miden el número de periodos de transferencia de in formación que tiene lugar cada segundo.

Los módems más antiguos empleaban técnicas de modulación como FSK y PSK, y por tanto en cada periodo de transferencia se enviaba un bit de información. Esto significa que la velocidad de transferencia toma el mismo valor, tanto en bps como en baudios (dicho de otro modo, se trabajaba a un bit por baudio).

Con el paso del tiempo, los módems alcanzaron el límite de baudios que las líneas telefónicas pueden ofrecer. Por ello, la única forma de aumentar la velocidad en bps ha consistido en utilizar nuevas técnicas de modulación, que empaquetan varios bits en cada periodo de transferencia de información (es decir, se trabaja a varios bits por baudio). Esto hace que la tasa de transferencia en bps sea mayor que su expresión en baudios. Por ejemplo, si se emplea modulación 8PSK, se envían tres bits en cada periodo de transferencia (3 bits por baudio). Esto implica que un módem 8PSK trabajando a 1.200 baudios ofrecerá una velocidad de transferencia de 9.600 bps.

En la actualidad, el concepto BBS ha perdido su interés, y la función principal del módem es facilitar el acceso a Internet. En este caso, el funcionamiento básico del módem sigue siendo el mismo. Nuestro PC utiliza el módem para marcar el número de un proveedor de servicios de Internet (denominado ISP, del inglés *Internet Service Provider*) . El ISP dispone de un banco de módems (en un número no siempre igualo mayor al número de usuarios suscritos a sus servicios, lo que puede provocar dificultades al conectar en horas "punta"), y uno de ellos atiende a la llamada de nuestro módem. Tras un protocolo inicial de negociación, el enlace entre nuestro PC y el ISP queda establecido, y el ISP

pasa a funcionar como intermediario entre nuestro PC y la Red. Más concretamente, el ISP forma parte de una red conectada a Internet, y al conectarnos, el ISP nos convierte en parte de dicha red.

Hoy día es posible encontrar módems con todo tipo de diseños: pequeñas tarjetas PCMCIA, tarjetas ISA, módems de sobremesa, etc. (Figura 1). Los módems externos presentan un conector que permite comunicar el módem con un puerto serie del PC. Además, suelen disponer de indicadores luminosos del estado de las líneas que controlan el módem, un zócalo para la conexión a la línea telefónica, y un segundo zócalo para conectar un teléfono, de forma que cuando el módem se desconecta, el teléfono se puede utilizar con normalidad.

Funcionamiento básico: técnicas de modulación

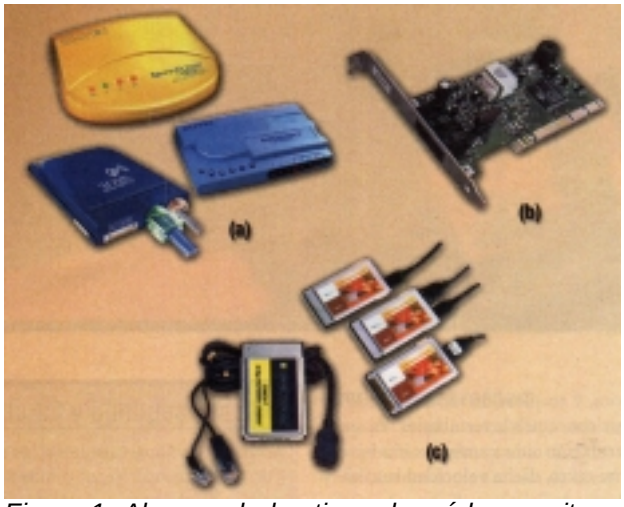


Figura 1. Algunos de los tipos de módems exitentes en el mercado: (a) módems externos; (b) módem interno; (c) módems en formato PCMCIA.

La palabra “módem” es una contracción de los términos “modulación» y “demodulación”. Parte de la base siguiente: se pretende enviar información digital a través de la red telefónica. La naturaleza de dicho medio permite enviar señales analógicas audibles de baja calidad (suficiente para el envío de voz, que es su propósito principal), pero no señales digitales. En el proceso de modulación, la información digital a enviar se convierte en una señal analógica audible que, ahora sí, se encuentra adaptada a la red telefónica, y por tanto puede ser transmitida. Cuando la señal llega al módem de destino, éste aplica el proceso de de- modulación, recuperando la señal digital original. Existen muchas técnicas de modulación de señales, tanto analógicas como digitales. Los objetivos principales de la modulación se centran en adaptar una señal al medio físico empleado, haciendo más sencilla su transmisión, y en permitir la transmisión simultánea de varias señales sin que exista interferencia (colocándolas en diferentes bandas de frecuencias). Por ejemplo, muchas emisoras de radio modulan sus señales analógicas en frecuencia (FM), facilitando su transmisión. En efecto, la señal resultante contiene frecuencias mucho más altas que las contenidas en la señal original, lo que permite la transmisión con antenas más pequeñas y empleando menor potencia. Además -tal y como hemos comentado antes- la modulación de las señales a distintas frecuencias permite que varias emisoras coexistan sin problemas. En este apartado presentaremos las técnicas de modulación más representativas del mundo de los módems que, en consecuencia, trabajan sobre señales digitales.

En los primeros módems (de 300 bps, el caso del Bell 103) se empleaba la modulación por desplazamiento de frecuencia, conocida como FSK (*Frequency Shift Keying*), una técnica derivada de la modulación en frecuencia (FM), pero

que trabaja con señales digita les. En el módem que efectúa la lla-mada, se definen dos tonos audibles distintos, es decir, dos señales con diferentes frecuencias (1.070 Hz y 1.270 Hz) y amplitud constante. Uno de los tonos (1.070 Hz) se asocia al estado lógico 0, mientras que el otro tono representa el estado lógico 1 (1.270 Hz). En el módem que ha recibido la llamada se hace lo mismo, definiendo dos frecuencias distintas de las anteriores (2.025 Hz para el 0 y 2.225 Hz para el 1). Esta diferencia de frecuencias permite que ambos módems usen la línea simultánea mente (la comunicación es *full duplex*). Los módems basados en modulación FSK alcanzan velocidades de transmisión muy reducidas, hasta 1.200 baudios (el término «baudio" se explica en el recuadro «Bits por segundo y baudios"). La Figura 2 ilustra el proceso de modulación FSK. Tratando de aumentar la velocidad en las comunicaciones, el siguiente paso se dirigió al empleo de la modulación por desplazamiento de fase (PSK, *Phase Shift Keying*). En este caso, las señales que representan los dos estados lógicos tienen una fase inicial diferente. En otras palabras, la señal correspondiente al cero digital inicia cada periodo con un nivel de tensión diferente al que caracteriza al 1 digital (ver Figura 3). Definiendo 4 fases diferentes en lugar de dos se consigue enviar un estado entre 4 posibles en cada transferencia, lo que equivale a 2 bits. Esta técnica se denomina 4PSK y permitió alcanzar los 2.400 bps a principios de los años 60. A finales de los 60 se introdujeron los módems basados en modulación 8PSK (8 fases diferentes), alcanzando los 4.800 bps.

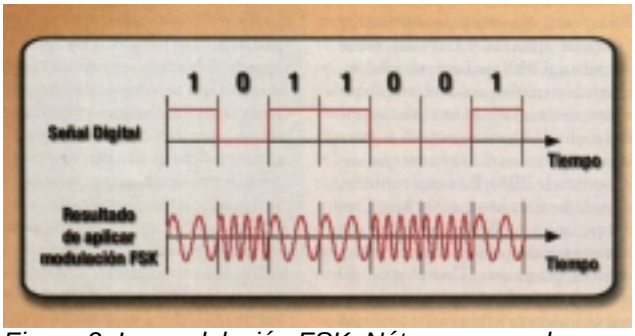


Figura 2. La modulación FSK. Nótese que cada estado digital queda representado por una frecuencia (alta para el estado "0" y baja para el estado "1").

Existe otra técnica llamada ASK, basada en la modulación en amplitud de señales analógicas (*Amplitude Modulation* o AM). Esta no se usa en los módems, pero conviene introducirla para comprender la siguiente sección de este artículo. En la modulación ASK se parte de una señal portadora con una determinada frecuencia, y se modifica su amplitud para diferenciar los estados 0 y 1. Generalmente, el estado lógico 0 se caracteriza por la ausencia de señal, mientras que el 1 se representa mediante la presencia de la señal portadora con una cierta amplitud.

La modulación QAM

La velocidad de transferencia se puede incrementar aún más gracias a la técnica de modulación de amplitud en cuadratura, conocida como QAM (*Qudrature Amplitude Modulation*). Esta técnica se aplica también en transmisiones de radio digital por microondas y en la transmisión de vídeo digital por cable (norma europea DVB-C). El objetivo clave es incrementar la velocidad de transferencia en bps, mientras se mantiene la misma velocidad en baudios. Es decir, en cada periodo de transmisión se empaquetan varios bits, en lugar de uno solo.

En QAM se parte de la combinación (suma) de dos señales caracterizadas por una misma frecuencia, pero desfasadas entre sí en un ángulo de 90° (se dice,

entonces, que dichas señales se encuentran en cuadratura). Este desfase significa, en términos más claros, que una señal está desplazada en el tiempo respecto de la otra en un cuarto del periodo que caracteriza a ambas señales. Una de las señales se denomina *In-Phase* (en fase, abreviada como "I"), mientras que la otra se denomina *Quadrature* (señal en cuadratura, abreviada "Q."). La amplitud de dichas señales se puede variar independientemente, siendo totalmente posible recuperar dichas amplitudes a partir de la señal combinada.

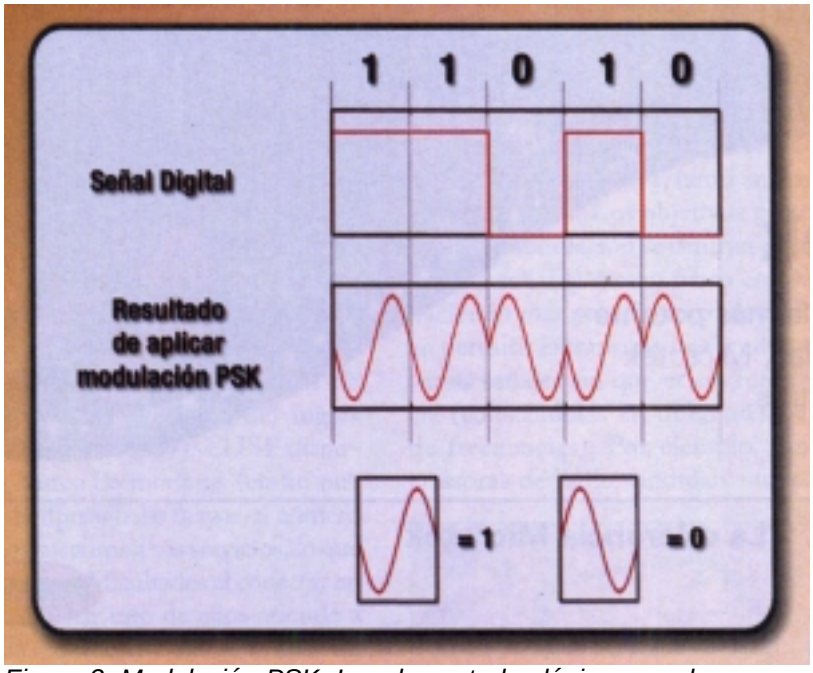


Figura 3. Modulación PSK. Los dos estados lógicos quedan representados por dos versiones de la misma señal, una desfasada respecto de la otra.

El proceso consiste en aplicar modulación en amplitud sobre la señal I, y sobre la Q, y finalmente sumarlas. De esta forma, se consiguen varios estados fase-amplitud, que se identifican mediante un conjunto de bits. Por ejemplo, la variante 8QAM define dos posibles amplitudes para cada una de las señales (sean A1 y A2), y transmite los siguientes estados amplitud-fase: A1-0°, A2-0°, A1-90°, A2-90°, A1-180°, A2-180°, A1-270° y A2 En total, hay 8 estados posibles, por lo que cada estado se puede codificar con 3 bits (ya que 23=8). Suponga que los estados expuestos arriba se codifican -respectivamente- como 000, 001, 010, 011, 100,101, 110 y 111. Si el módem recibe A1-0°, A2-0°,A1-180° y A2- 90°, la demodulación interpretará la pa labra digital siguiente: 000 001 100 011. Por tanto, con tan sólo 4 periodos de transmisión, se pueden transmitir 12 bits. Es decir, la velocidad en bps es el triple de la velocidad en baudios.

Otra variante de la modulación QAM es 16QAM, que define 4 posibles amplitudes para la señal 1, y otras 4 para la señal Q resultando en un total de 16 posibles estados. Esto significa que se enviarán 4 bits en cada periodo de transmisión. Empleando 16QAM aparecieron, a finales de los años 60, los módems a 9.600 bps.

También existe (fuera del contexto de los módems) la variante 32QAM. En es te caso, hay 6 estados definidos para la componente I y otros 6 para la componente Q, lo que resulta en 36 estados posibles. Cuatro de ellos se descartan, con lo que quedan 32 estados posibles (5 bits por periodo de transmisión). Incluso se llega a la variante 256QAM (también fuera del contexto del módem), que envía paquetes de 8 bits. Esta variante no resulta demasiado eficiente, puesto que tal cantidad de estados hace que la separación entre ellos (en tér-

minos de amplitudes y fases) sea muy reducida. Esto motiva que las señales de ruido puedan transformar un estado en otro con mayor facilidad que en el resto de variantes. La solución radica en transmitir la señal con mayor potencia, pero esto reduce la eficiencia con respecto, a las anteriores soluciones. Se puede encontrar una descripción detallada de la modulación QAM en http://www.ece.wpi.edu/courses/ee535/ hwk97/hwk3cd97/mrosner/node5.html

La Figura 4 muestra el aspecto de una señal modulada mediante esta técnica.

Las técnicas de modulación fueron evolucionando, llevando la velocidad de transferencia a 14,4 kbps e incluso 28,8 kbps. Todavía se consiguió llegar, en la década de los 90, a 33,6 kbps, lo cual ya se consideraba un limite difícil de superar por la propia naturaleza de la red telefónica.

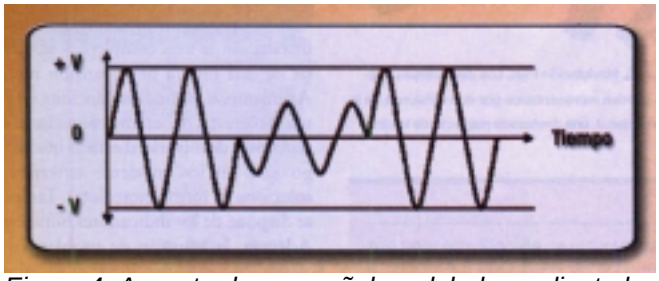


Figura 4. Aspecto de una señal modulada mediante la técnica QAM.

Módems a 56k

Para sorpresa de muchos, en febrero de 1998 se introdujo el estándar V.90, que definía módems capaces de alcanzar los 56 kbps (comúnmente llamados 56k), llegando mucho más allá de los límites teóricos que la red telefónica puede alcanzar. Estos son los módems telefónicos más rápidos que se pueden adquirir hoy día. Fueron introducidos por las firmas US Robotics y Rockwell.

En primer lugar, hay que aclarar que la velocidad de 56 kbps no es una constante: es un límite máximo que se alcanza cuanto la red telefónica se encuentra en estado óptimo. De hecho, rara vez se llega a ese límite, alcanzando velocidades máximas (en los mejores casos) de unos 52 kbps.

Por otro lado, hay que tener en cuenta que el ISP debe poseer módems adecuados al otro extremo para gozar del incremento de velocidad (si dispone de módems de 33.6 kbps, ésa será la velocidad máxima alcanzable).

Otro factor a considerar es la propia línea telefónica empleada. Si usando un módem de 33.6 kbps la línea no permite alcanzar dicha velocidad, resulta evidente que nada se va a ganar con un módem de 56 k. En resumen, el hecho más importante a recordar es el siguiente: la línea telefónica es la que condiciona la velocidad que se puede alcanzar. Por tanto, los usuarios conectados a su ISP mediante una línea de baja calidad no harán una buena inversión al adquirir un módem de 56 k

El estándar de fabricación de módems más reciente es el V.92. Sigue apuntan do a los módems 56 k, pero introduce ciertas ventajas. En primer lugar, cuando se recibe una llamada mientras estamos conectados, el módem suspende la sesión y, tras finalizar dicha llamada, reactiva la sesión sin necesidad de repetir el proceso de conexión con el ISP. Esta mejora también funciona a la inversa: cuando se realizan llamadas, la conexión se desactiva, y se reanuda rápidamente al finalizar la llamada. Por supuesto, mientras se atienden o se realizan

llamadas, la actividad de la conexión al ISP queda suspendida (y por tanto no se pueden realizar descargas, navegar, etc.). Para que esta característica funcione, es necesario tener activado el servicio de llamada en espera, ofrecido por la mayoría de las compañías telefónicas actuales.

La norma V.92 también ofrece la característica *quick-connect* (conexión rápida). Esta técnica evita ciertas fases de la etapa de negociación que ocurre entre el módem del PC, y e! módem del ISP. Para hacer esto, se utilizan ciertos detalles de la línea telefónica, que fueron memorizados en la última sesión. Esta técnica puede llegar a incrementar la velocidad de conexión en un 50%.

Finalmente, es posible incrementar la velocidad de envío de datos del usuario al ISP. Generalmente, los datos entrantes (descargas, por ejemplo) pueden llegar a mayor velocidad que los datos salientes (e mails, ficheros enviados a un servidor FTP, etc.). Esto parece lógico, ya que la mayoría de usuarios descargan más información que la que envían (páginas web, ficheros, etc.). Con la norma V.92, ahora es posible aumentar, incluso hasta el máximo, la velocidad de envío de datos (sacrificando la velocidad de descarga). Esto supone una ventaja importante para los usuarios que envían grandes cantidades de información (gran volumen de e-mails, fotografías digitales, etc.).

Tal y como ocurría con la norma V.90, es necesario que el ISP se encuentre adaptado al estándar V.92 para que podamos gozar de las características arriba expuestas.

Módems internos y externos

A la hora de adquirir un módem, no tardamos en percibir dos posibles variantes. Una de ellas la forman los módems externos (Figura 1-a), el clásico periférico externo que integra indicadores luminosos para mostrar el estado de las líneas, un interruptor de encendido! apagado, un conector serie, una entrada de línea telefónica, un cable de alimentación y, habitualmente, una entrada para conectar un teléfono.

Por otro lado, tenemos la oportunidad de adquirir un módem implementado como una tarjeta de expansión (Figura 1- b), que se conecta a una ranura libre en la placa base, y después se configura siguiendo los procedimientos habituales. Ante tal elección, hay que tener en cuenta que todo apunta como mejor solución al uso de los clásicos módems externos. La ventaja más importante radica en que el módem trabaja de forma totalmente ex terna e independiente del PC, implementando toda la funcionalidad dentro de una caja.

En el caso de los módems internos, la instalación es más compleja, y se dispone de una ranura de expansión menos. Asimismo, si el módem funciona de forma incorrecta, deberemos reiniciar el sistema para devolverlo al estado inicial (algo que en los módems externos se soluciona de forma inmediata). Tampoco se dispone de los indicadores luminosos. Además, la solución de problemas requiere, en ocasiones, llevar el PC al servicio técnico. Como ventajas, el precio es menor, el cableado necesario es mínimo, y no ocupa espacio externo al PC.

En e! mundo del PC portátil se emplean con frecuencia los módems PC Card, que presentan un tamaño similar al de una tarjeta de crédito (Figura 1-c). Estos funcionan de forma interna, pero se insertan y extraen con facilidad, y constituyen una solución apropiada para los usuarios que viajan con frecuencia.

Todavía existe otra categoría: los llamados WinMódems, que se implementan como módems internos pero están controlados mediante software. Este tipo de módems se describe en el recuadro 'WinMódems".

WinModems

Bajo los conocidos términos *WinModem* y *Software Modem* se esconde una idea simple. Se trata de extraer varias funciones implementadas por el hardware del módem, e implementarlas mediante software ejecutado en el PC en forma de *drivers.* Esto simplifica el diseño considerablemente ya que el hardware a fabricar se reduce en gran medida. El tamaño del modem y su precio son menores. El consumo eléctrico también se reduce. Además, esta aproximación permite actualizar el módem de una forma realmente sencilla (evitando actualizaciones del *firmware* del modem) tan solo hay que descargar la ultima versión de los correspondientes *drivers* y proceder con la instalación. Tras la aparición de este tipo de módems los usuarios reportaban gran cantidad de problemas respecto a su funcionamiento. Actualmente los *WinModems* han superado estos problemas principalmente debido al aumento del rendimiento de los PC (la implementación del modem mediante software supone una carga de computo extra para la CPU). No hay que olvidar que este tipo de modem depende totalmente del sistema operativo empleado ya que los *drivers* se desarrollan para un sistema operativo en particular (prácticamente se puede decir que todos funcionan exclusivamente bajo Windows y de ahí el nombre) Por ello en caso de no emplear Windows se debe recurrir a un módem convencional (implementado por hardware) y no a un *WinModem* Como dato practico los *WinModems* suelen incluir una indicación en su caja recordando que requieren Windows para funcionar. A la hora de adquirir un WinModem es recomendable detenerse a estudiar si el fabricante proporciona suficiente soporte en línea y si actualiza los drivers con frecuencia. También hay que tener en cuenta que algunos *WinModems* no se adaptan a todas las versiones del sistema operativo Si se planea una actualización del sistema operativo, es muy recomendable cerciorarse de que los drivers oportunos están disponibles. También es conveniente recoger información de los usuarios ya que algunos drivers son inestables y como consecuencia hacen que Windows sea inestable

También es cierto que algunos WinModems son tan baratos que muchos usuarios adquieren uno nuevo al actualizar el sistema operativo

CAPÍTULO 17

LAS IMPRESORAS

En esta entrega presentamos el funcionamiento de un periférico clásico pero imprescindible: la impresora. Ésta permite transformar los textos y gráficos electrónicos en auténticos documentos impresos.

En la actualidad, cualquier usuario de PC trabaja habitualmente en formato electrónico. Se redactan documentos mediante herramientas de proceso de texto, se crean presentaciones que se pueden mostrar directamente en la pan talla del PC, se escriben “cartas electrónicas” mediante e-mail, etc. Sin duda, el uso innecesario del papel se ha reducido considerablemente, lo que conlleva muchos beneficios (por ejemplo, el consiguiente impacto positivo en el ámbito ecológico o la reducción de costos debido al menor gasto de papel).

Sin embargo, se hace necesario el paso a versión impresa en un punto u otro de la vida de la mayoría de los documentos. La salida mediante un monitor no persiste durante años, y tampoco es transportable, mientras que la versión en papel sí que cumple con estas características. El periférico capaz de transformar texto y gráficos desde su versión digital hacia su forma impresa recibe el nombre de impresora, y constituye un dispositivo esencial para cualquier usuario de PC. De hecho, mucho antes de la aparición del PC, las impresoras ya se empleaban con los ordenadores primitivos como me dio principal para la presentación de resultados. Y de hecho, todo parece apuntar a que las impresoras no se abandonarán en el futuro.

En este artículo presentaremos el funcionamiento básico de las impresoras y las tecnologías de impresión más aceptadas actualmente. En particular, prestaremos atención a las dos tecnologías más extendidas: las impresoras láser y las de inyección de tinta.

Introducción a las impresoras

Para llevar a cabo el proceso de impresión, toda impresora consta de tres subsistemas básicos: hardware de control, sistema de transporte del papel y un mecanismo de impresión sobre el papel. El hardware de control se encarga de gobernar el funcionamiento de los componentes de la impresora. El mecanismo de impresión hace que los caracteres y gráficos a imprimir queden efectivamente “dibujados” sobre el papel. Suele consistir en un cabezal de impresión que se puede desplazar horizontalmente. Finalmente, el sistema de transporte desplaza el papel verticalmente, haciendo que la tinta va ya a parar, finalmente, al lugar oportuno en el papel (es decir, a la línea oportuna). El origen de la información a imprimir suele adoptar tres formatos básicos: texto (secuencias de códigos ASCII), objetos definidos vectorialmente (es decir, matemáticamente) o bien mapas de bits o bitmaps (que definen todo elemento a imprimir como un conjunto de puntos). En general, y al igual que ocurría con los monitores, las impresoras forman las imágenes y el texto a partir de puntos (píxeles). Las im-

presoras suelen estar dotadas de una memoria ROM, que alma cena el mapa de bits (bitmap) correspondiente a cada carácter, e incluso una memoria RAM que permite que el PC envíe otras fuentes a la impresora.

Características básicas

La caracterización de las impresoras en cuanto a prestaciones se lleva a cabo mediante cuatro parámetros fundamentales. En primer lugar, la velocidad de la impresora se determina en páginas por minuto (ppm) o bien en caracteres por segundo (cps). En la actualidad, se usa prácticamente siempre la unidad ppm, y se reserva la velocidad en cps para las impresoras matriciales (muy poco extendidas en comparación con el resto de tipos). A la hora de interpretar la velocidad especificada por el fabricante, debemos ser realmente cautos, e indagar en los detalles: ¿cómo se ha medido dicha velocidad? Normalmente los fabricantes indican que su impresora alcanza 6 páginas por minuto, pero no especifican que se trata de páginas con un 5% de información impresa, sin gráficos y en baja calidad. Incluso se suele descontar el tiempo de cálculo empleado por el ordenador, aumentando más la cifra. Esta cifra es la máxima que puede alcanzar el motor de la impresora.

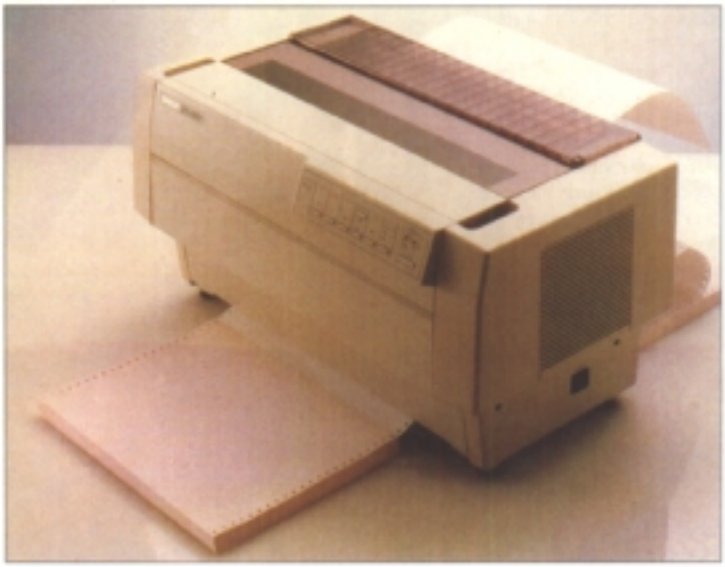


Figura 1. Aspecto de una impresora de matriz de puntos (Epson DFX-5000+)

La *resolución* de la impresora es un parámetro íntimamente ligado a la calidad de impresión. Indica la cantidad de puntos (píxeles) que la impresora puede crear sobre el papel, por unidad de superficie. Se suele medir en puntos por pulgada (ppp), tanto en dirección horizontal como vertical. Por ejemplo, una impresora con resolución de 600 x 300 ppp es capaz de imprimir 600 puntos en cada 2,54 cm. hori-

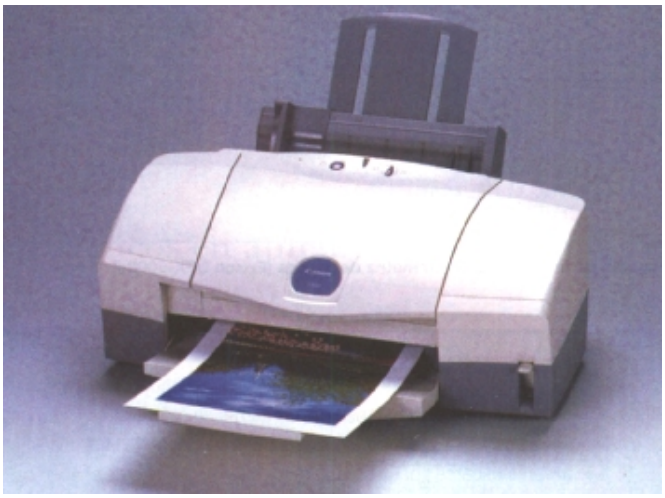


Figura 2. Ejemplo de impresorade inyección de tinta tipo bubble jet (Canon S800)

zontales (una pulgada), y 300 puntos en cada pulgada vertical. Si sólo se indica un número, la resolución es la misma en ambas direcciones (por ejemplo, 600 ppp equivale a 600 x 600 ppp). No hay que olvidar que la resolución no es directamente traducible en calidad. Si la impresora presenta una elevada resolución, pero no sitúa los puntos con precisión sobre el papel o los puntos son demasiado gruesos, el resultado no presentará alta calidad.

El tamaño del buffer de memoria (zona de almacenamiento temporal de da tos en la impresora) es otro dato importante, ya que determina el rendimiento de las comunicaciones entre el PC y la impresora. El PC funciona a una velocidad considera-blemente más rápida que la impresora. Por tanto, sin un buffer, el PC debería esperar continuamente a la impresora entre envío y envío. Gracias al buffer, el PC envía datos a la impresora, y pasa a realizar otras tareas mientras la impresora procesa dicha información.

A mayor tamaño de buffer, más rápida es la impresión. El tamaño habitual es de 256 kB, aunque las impresoras más profesionales ofrecen hasta varios MB.

Figura 3. Cartuchos tipo bubble jet (Canon S600)

Finalmente, el último parámetro de interés es la interfaz de conexión. Hasta hace poco la más habitual era el puerto paralelo estándar del PC, utilizando el conector centronics de 36 terminales (ver entrega de esta serie en PC World n° 188, de junio de 2002). También existen impresoras que funcionan a través del puerto serie RS-232, lo que minimiza el número de cables a utilizar y permite emplear cables mucho más largos. Sin embargo, la impresión serie resulta mucho más lenta, por lo que no es la interfaz de conexión más habitual. Hoy en día, la conexión vía USB es la más común por su elevada velocidad frente al puerto paralelo.

Otras conexiones habituales, normalmente compartidas con una de las anteriores, son los puertos de infrarrojos, de red o hasta un enlace Bluetooth inalámbrico.

Tecnologías básicas de impresión

Existen dos tecnologías básicas de impresión: las que se basan en impacto (matriz de puntos y caracteres) y las que trabajan sin impacto. Las impresoras basadas en una matriz de puntos (Figura 1) con tienen un grupo de “agujas” que se asientan sobre un cabezal móvil. Estas agujas impactan sobre una cinta impregnada de tinta (mediante la aplicación de fuerza producida por electroimanes), lo que ha ce que la tinta se transfiera al papel en cada pequeño punto de impacto. Estas impresoras eran muy populares antes de la aparición de las

impresoras de inyección de tinta. Hoy en día, aún se utilizan en algunos contextos, debido a su capacidad de usar papel autocopiativo. Como des ventaja, hay que resaltar que se trata de dispositivos lentos y con baja calidad de impresión.

Las impresoras de caracteres son, básicamente, máquinas de escribir compute- rizadas. Contienen una serie de barras con la forma de cada carácter predefinida. Para imprimir un carácter, la barra correspondiente mueve el patrón del carácter con fuerza hacia la cinta impregnada de tinta, y por tanto el carácter se transfiere al papel. En este caso, cada carácter se transfiere como un todo (mediante un único impacto), y no se forma por un conjunto de impactos (puntos). Las impresoras de caracteres son muy rápidas para la impresión de texto, pero no es posible incluir gráficos.

Figura 4. Impresora de inyección de tinta HP DeskJet 920c (Hewlwtt-Packard).

Las impresoras sin impacto producen las imágenes y el texto sin llegar a tocar el papel. En otras palabras, utilizan técnicas que permiten guiar la tinta hacia el papel, colocándola en el lugar preciso y en cantidad oportuna. Dentro de esta tecnología, destacan dos tipos básicos de impresora: inyección de tinta y láser. Estos tipos de impresora son los más aceptados actualmente, y los comentamos en detalle en los siguientes apartados. Además, existen otras muchas tecnologías de impresión sin impacto, que se comentan en un recuadro independiente de este artículo.

Impresoras de inyección de tinta

Desde su introducción en la segunda mitad de los años 80, las impresoras de inyección de tinta (*ink-jet* en inglés) han gozado de una aceptación en constante crecimiento. Su precio no ha parado de descender y sus prestaciones no han dejado de aumentar. Pero, ¿qué es una impresora de inyección de tinta?

En este tipo de impresoras, un cabezal de impresión dispara pequeñísimas gotas de tinta (con un diámetro de unas 50 micras, menor que el diámetro de un cabello) sobre el papel, formando finalmente texto e imágenes. El cabezal (que transporta los cartuchos de tinta) se desplaza en sentido horizontal, mientras que la hoja se desplaza línea a línea en sentido vertical, gracias a un mecanismo de transporte del papel. Las gotas de tinta se posicionan sobre el papel con una precisión extrema, alcanzando resoluciones de hasta 4.800 ppp. Ya que cada punto puede tener un color diferente, se pueden generar imágenes impresas de calidad fotográfica.

Dentro de este tipo de impresoras se distinguen dos tecnologías básicas: *bubble jet* (introducida por Canon) y *desk-jet* (introducida por Hewlett-Packard), donde la diferencia reside, básicamente, en el modo de generar las gotas de tinta.

En las impresoras *bubble jet* o de inyección térmica, se aplica calor sobre la tinta, que se halla situada en un depósito dentro del cartucho de impresión, del que fluyen varios micro-conductos por los que saldrá la tinta. Esto se consigue haciendo pasar un impulso de corriente eléctrica a través de unas resistencias. El calor hace que la tinta entre en estado de ebullición, generando una burbuja que crece en volumen, y empuja a la tinta hacia el exterior, a través de los conductos. Este proceso dura aproximadamente un milisegundo, y desaloja un volumen de tinta predeterminado (una gota). La presión de la burbuja produce un efecto “cañón”, que dispara la gota sobre el papel. Cada vez que la corriente en las resistencias cesa, la burbuja desaparece, y por tanto se produce un efecto de succión que toma tinta del depósito y rellena los conductos.

En el caso de las impresoras *desk-jet* se emplean cristales piezoeléctricos como elemento fundamental, en lugar de resistencias. Se aprovecha la característica básica de un cristal piezoeléctrico: si se aplica tensión eléctrica, se produce una de formación del cristal. Por tanto, se envían los impulsos eléctricos a los cristales, y su deformación produce un bombeo de la tinta desde el depósito hacia los micro-conductos, disparando la tinta hacia el papel. Esta filosofía de funcionamiento es similar a la de un gotero.

Los cartuchos más habituales suelen contener tinta en estado líquido, por lo que ésta no necesita ningún tratamiento previo a la impresión. Sin embargo, existen cartuchos en los que, a temperatura ambiente, la tinta se encuentra en estado sólido. En este caso, se utilizan resistencias para pasar la tinta a estado líquido antes de ser disparada hacia el papel. Durante el recorrido hacia el papel, la tinta se va solidificando, y queda finalmente adherida al papel sin ser absorbida. Esto evita un problema típico de la tinta común: la imagen impresa se encuentra “seca”, y no es necesario utilizar papel especial para evitar este efecto. Los cartuchos correspondientes a las impresoras *deskjet* suelen ser más baratos, ya que tan sólo contienen el cartucho en sí, y no el cabezal de impresión completo (algo que ocurre en los cartuchos *bubble jet*).

El papel a imprimir se carga positivamente en su totalidad. Por tanto, al hacerlo pasar por el tambor, atraerá a las partículas de tóner (que tienen carga negativa), y la imagen quedará finalmente formada sobre papel. Finalmente, el tóner adherido al papel se funde mediante la aplicación de calor, haciendo que quede totalmente fijado al papel. Se consigue así imprimir una página en una sola pasada, al contrario que en las impresoras de inyección de tinta, donde la página se imprime línea a línea. Antes de imprimir una nueva página, se realiza un borrado electrostático del tambor, dejándolo preparado para un nuevo ciclo.

Si planea imprimir gran cantidad de documentos con mucha frecuencia, la elección más rentable es una impresora láser. Sin embargo, si desea imprimir documentos de tamaño medio con una frecuencia moderada, elija una impresora de inyección de tinta

Impresoras láser

La impresión láser se basa enteramente en la interacción electrostática, el mismo fenómeno que produce que un plástico atraiga trozos de papel tras ser frotado con una prenda de fibra. Para comprender la impresión electrostática, basta saber que las cargas eléctricas pueden ser positivas o negativas, y que las cargas de signo opuesto se atraen, mientras que las cargas de igual signo se repelen.

En primer lugar, se carga negativamente toda la superficie de un tambor fotosensible, del tamaño de una hoja. Acto seguido, se hace avanzar el tambor línea a línea, y un láser recorre horizontalmente cada línea, ayudado por un espejo giratorio (en otras palabras, se produce un proceso de barrido). El láser incide en los puntos donde la tinta se deberá fijar, invirtiendo la carga (que ahora será positiva). El láser se desconecta en los lugares donde no deberá aparecer tinta (quedan do con carga negativa). Por tanto, tras re correr todo el tambor, solo habrá cargas positivas en los puntos donde deberá depositarse tinta, mientras que el resto (lo que constituirá el fondo blanco del papel) queda cargado negativamente. En otras palabras, se ha conseguido crear una imagen electrostática de la hoja a imprimir, mediante cargas positivas sobre un fondo de cargas negativas.

Figura 5. Impresora láser (HP LaserJet 9000)

Los puntos cargados positivamente en el tambor atraen partículas de tóner (mate-rial electronegativo mezclado con un pigmento que lo dota de color). Por tanto, la imagen final queda “dibujada” sobre el tambor por medio de puntos negros de tóner.

Existe otra variante de las impresoras láser en las que no es necesario un pro ceso de barrido. En lugar de un láser y un sistema de espejos se dispone de una hilera de diodos emisores de luz (Láser LED). Por ejemplo, en una impresora de 300 ppp, habrá una hilera de LED cubriendo una línea completa del papel, a razón de 300 LED por pulgada. Sólo se encienden, para cada línea, aquellos dio dos que corresponden a puntos donde deberá aplicarse tóner. Este proceso se repite línea a línea hasta procesar el tambor completo. Se produce el mismo efecto que con un barrido láser, pero de forma más rápida.

Otra variante emplea diodos de cristal líquido (LCD) en lugar de LED. Estos conforman un material que es transparente u opaco según el nivel de tensión eléctrica que se le aplica. Se forzarán al estado transparente aquellos cristales correspondientes a los puntos donde deba aplicarse tóner, manteniendo el resto de diodos en estado opaco. Por otra parte, se aplica una lámpara halógena que ilumina todos los cristales, y sólo pasa luz a través de los diodos en estado transparente, invirtiendo la carga en el tambor.

Las impresoras láser (Figura 5) son mucho más rápidas que las impresoras de inyección de tinta. Además, están dotadas de una mayor precisión en la colocación de puntos sobre el papel. También economizan tinta, ya que depositan la cantidad de tóner necesaria, sin exceder ese límite. El tóner (Figura 6) no es caro en comparación con los cartuchos de tinta y, además, es mucho más duradero, lo que resulta rentable en el entorno de una oficina, donde se imprimen gran cantidad de documentos diariamente. Como desventaja principal, el precio de estas impresoras es muy elevado en comparación con las impresoras de inyección de tinta. En conclusión, si planea imprimir gran cantidad de documentos con mucha frecuencia, la elección más rentable es una impresora láser. Sin embargo, si desea imprimir documentos de tamaño moderado con una frecuencia también moderada, la mejor elección es una impresora de inyección de tinta.

Figura 6. Módulos de tóner de una láser color 8Lexmark OptraColor 1200)

Formación del color

En la entrega de esta serie dedicada a los monitores dijimos que el color se forma a partir de la combinación de tres colores primarios (rojo, verde y azul). En el caso de las impresoras, los colores primarios utilizados son justamente los complementarios del sistema RGB: cián (complementario del rojo), amarillo (complementario del azul) y magenta (complementario del verde). Si estos tres colores se mezclan entre sí, el resultado debería ser el color negro. Sin embargo, ya que estos colores no son puros, se obtiene un tono “café”. Por ello, a este sistema se le añade el negro, formando el sistema denominado CYMK (C para cián, Y para amarillo, M para magenta y K para negro). En conclusión, una impresora sólo contiene cuatro tintas distintas.

A la hora de imprimir un punto, la impresora lo forma como un “superpunto”, formado por un entrelazado de puntos de los distintos colores (CYMK). El ojo (por su menor resolución espacial y por la distancia al papel) tiende a realizar una combinación de los colores, haciendo que veamos el superpunto con un color determinado.

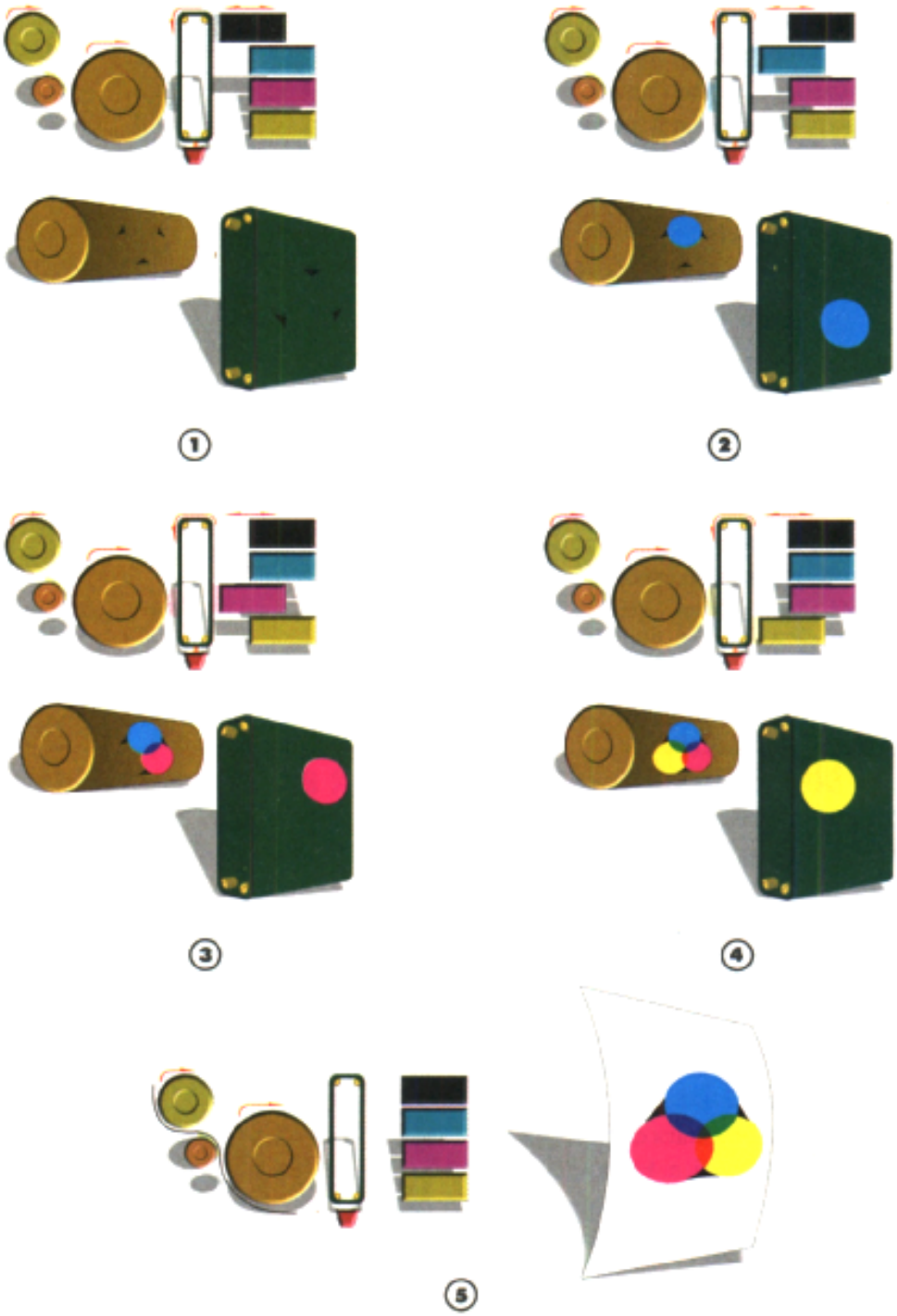


Figura 7. En una impresora láser color, primero se aplica el tóner negro sobre el fotoconductor y se pasa al tambor; después se hace lo mismo con el cián, magenta y amarillo. Cuando la imagen completa está formada en el tambor, se pasa al papel y se fija fundiendo el tóner.

Otras tecnologías de impresión sin impacto

Además de las conocidas impresoras láser y de inyección de tinta, existen otras tecnologías que permiten crear documentos impresos evitando el contacto con el papel. En primer lugar, las impresoras de sublimación de tinta ofrecen la posibilidad de imprimir imágenes de altísima calidad. Se basan en el empleo de una cinta transferible formada por una película plástica. Sobre la cinta se sitúan paneles del tamaño de una página, cada uno recubierto con tinta (en estado sólido) de un color: cián, magenta, amarillo y negro (CYMK). El cabezal de impresión contiene miles de elementos genera dores de calor, que son capaces de controlar su temperatura con precisión extrema. El cabezal se desplaza a través de la cinta, y el calor aplicado por los elementos calentadores hace que la tinta se vaporice, di fundiéndose sobre la superficie del papel (el proceso de cambio de estado sólido a gaseoso se denomina sublimación, y de ahí el nombre de esta tecnología). El cabezal realiza una pasada completa sobre la página para cada color básico, construyendo la imagen de forma gradual. Los diferentes colores se obtienen gracias a las variaciones de calor: a mayor temperatura, más tinta se difunde y más intenso es cada color. La ventaja indiscutible de este tipo de impresoras es la excelente calidad de imagen que es capaz de producir. Las impresoras de sublimación de tinta son utilizadas ampliamente por servicios de publicaciones, artistas gráficos y fotógrafos profesionales. Sin embargo, estas impresoras no resultan apropiadas para la impresión de documentos cotidianos", ya que el coste de impresión es elevado y la impresión es lenta (requiere varias pasadas).

Las impresoras de tinta sólida (un sistema muy ligado a la firma Tektronix, ahora Xerox) contienen cuatro barras de color (CYMK), formadas por tinta en estado sólido. La tinta se derrite mediante la aplicación de calor, y se esparce por un tambor de transferencia, que es el que finalmente imprime cada página de una sola pasada. Estas impresoras se caracterizan por un bajo coste de adquisición y mantenimiento, y por una gran calidad de impresión en prácticamente cualquier tipo de papel. Por ello se utilizan con frecuencia para preparar transparencias e impresiones de gran des dimensiones.

Por otra parte, las impresoras de cera térmica contienen una cinta formada por paneles del tamaño de una página, correspondiendo a los cuatro colores básicos (CYMK). Al imprimir, la cinta pasa a través de un cabezal de impresión térmico. Éste contiene miles de finas agujas de impresión, capaces de controlar la temperatura con elevadísima precisión. La cera se funde y se deposita sobre un papel dotado de un revestimiento especial o sobre una transparencia. La imagen final está compuesta de minúsculos puntos de cera de color. Estas impresoras ofrecen un bajo coste por página, y una rapidez aceptable. Sin embargo, requieren del uso de papel especial y su calidad no supera a la ofrecida por las impresoras de sublimación.

Finalmente, cabe citar la tecnología de impresión térmica autócroma. En este caso, el color se encuentra en el papel, y no en la impresora. El papel contiene tres capas de color: cián, magenta y amarillo. Cada capa se activa mediante la aplicación de una cierta cantidad de calor. El cabezal de impresión realiza tres pasadas (una por color), aplicando la cantidad de calor oportuna para activar cada capa.

Impresoras GDI *(Win Printers)*

Las impresoras GDI o *Win Pririters* se basan en una tecnología propia de Windows, llamada GDI *(Graphical Device Interface).* GDI es una librería que permite desarrollar impresoras que dejan gran parte del trabajo de impresión al sistema operativo Windows. Por ejemplo, la impresora puede prescindir de memoria, empleando la RAM del PC. Todo esto deriva en una reducción de la cantidad de hardware a implementar en la impresora, ya que parte de él se implementa mediante software. Además, como era de esperar, dicha reducción hace que el costo de la impresora sea considerablemente menor (se pueden llegar a ahorrar unos 60 euros).

Al igual que ocurría con los Win-modems, las impresoras GDI sólo funcionan bajo el sistema operativo Windows. Si se planea migrar a otro sistema, no se debería adquirir una impresora de este tipo. Conviene tener en cuenta que, en principio, puede ocurrir que la impresora no funcione con futuras versiones de Windows. Es muy importante asegurarse de que los controladores de la impresora se encuentran bien documentados y el soporte que proporciona el fabricante sea de calidad.

Ya que la impresora funciona mediante Windows, el sistema operativo se verá obligado a cargar con más tareas, lo que puede volver el sistema lento o inestable si su PC no es suficientemente potente.

CAPÍTULO 18

EL RATÓN Y EL TECLADO

En este capítulo abordamos dos componentes de entrada directamente relacionados con la interacción entre el usuario y el PC: el ratón y el teclado. Se trata de dos dispositivos con una función bastante específica, económicos y con un diseño no excesivamente complejo. Quizá por ello acostumbramos a prestarles poca atención. Sin embargo, son componentes imprescindibles para interactuar con un PC, y de hecho son los que más utilizamos. Piense en cualquier sesión de trabajo con su PC empleando Windows: pasa la mayor parte del tiempo utilizando el teclado y el ratón. Por ello, resulta conveniente echar un vistazo a su interior y comprender su funcionamiento, objetivo principal de este artículo.

Introducción al ratón

Resulta asombroso que -a pesar de desarrollarse en los años 60- un componente tan sencillo y necesario como el ratón tardara tanto tiempo en aparecer en el mundo de los ordenadores (años 80). Con frecuencia, el ser humano encuentra necesaria la acción de señalar durante cualquier acto de comunicación. En consecuencia, en la interacción con una computadora, dicha necesidad sigue ahí. Por esto sorprende la tardía aparición del ratón.

Hay que decir que, en los inicios de la computación, el ratón no tenía razón de existir En efecto, en aquellas primitivas computadoras, la Interfaz hombre-máquina era también primitiva (tarjetas perforadas, etc.). En los años 60 y 70, la interfaz de usuario se basaba en texto. El usuario disponía de la ayuda de las teclas denominadas "cursores", que permitían desplazarse a tra-vés de la interfaz de usuario en programas como los editores de texto, etc. Era una primera solución para hacer posible apuntar o señalar. En los años 70 se hicieron populares dispositivos como los lápices ópticos, tabletas gráficas *joystick*, que ofrecían métodos más avanzados.

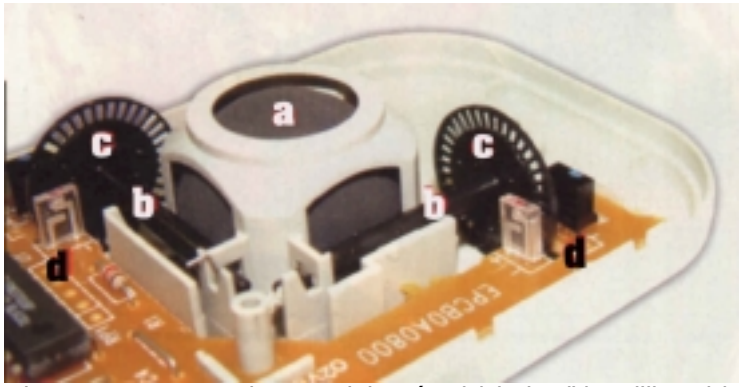


Figura 1. Estructura interna del ratón: (a) bola, (b) rodillos, (c) Disco perforado, (d) LED infrarrojos.

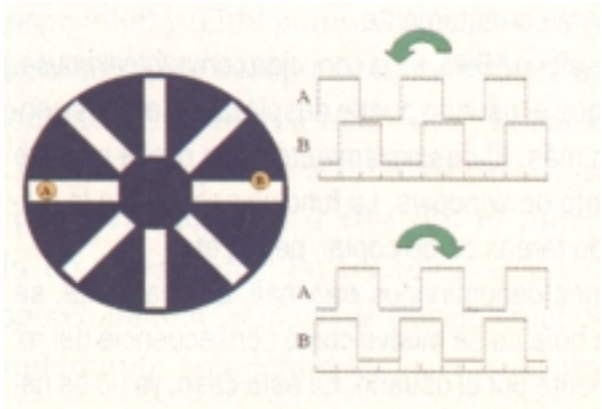


Figura 2. Detección del sentido de desplazamiento. En la parte derecha se aprecia el desfase entre las señales detectadas por ambos sensores, que determina el sentido de giro

Fue en 1973 con el sistema Alto de Xerox cuando se utilizó el primer ratón (que se había presentado en 1968) para sacar provecho a la primera interfaz gráfica. En 1984 -con la introducción del ordenador personal Macintosh- el ratón empezó a popularizarse, alcanzando un éxito rotundo y redefiniendo completamente la forma en que empleamos los ordenadores. El ratón se presenta como un dispositivo simple, pero que ofrece una forma de apuntamiento realmente eficiente. Las aplicaciones actuales están tan sumamente orientadas al uso del ratón que, en su ausencia, el manejo se convierte en lento y pesado. Sin duda, hoy en día un elevado porcentaje de las acciones que realizamos con un PC se desencadenan mediante unos simples movimientos y pulsaciones realizados con el ratón.

En el mundo del PC, el ratón tardó un poco más en aparecer. El estallido ocurrió cuando Windows 3.1 ganó popularidad y las interfaces gráficas de usuario se convirtieron en un estándar. En la actualidad, todo usuario de PC pasa horas usando el ratón para apuntar a objetos presentados en pantalla, con la intención de activarlos, arrastrarlos, soltarlos, redimensionarlos, etc. Veamos cómo funciona el ratón.

Funcionamiento del ratón

La misión principal del ratón consiste en señalar puntos concretos de la interfaz de usuario de los programas. Esto se traduce en convertir los movimientos de la mano -deslizando el ratón sobre una superficie plana- en información digital que el ordenador

Teclados ergonómicos

Los conocidos teclados ergonómicos tienen como objetivo proporcionar un medio cómodo para teclear, haciendo que manos, muñecas y antebrazos se coloquen en una posición más relajada, con respecto a los teclados convencionales. Algunos estudios revelan que el uso del teclado en un modo inapropiado puede derivar en lesiones como la tendinitis y el síndrome discarpial.

El teclado queda dividido en dos grupos de teclas, que se disponen formando un cierto ángulo. De esta manera, los codos reposan en una posición mucho más natural que la usual. También se suele añadir un reposamuñecas y se aplica una cierta curvatura al teclado. Entre los teclados ergonómicos disponibles en el mercado, cabe destacar el producto Natural Keyboard de Microsoft (Figura 6). Hay que remarcar que el uso de estos teclados implica un cierto periodo de familiarización con la nueva organización de teclas. En general, el usuario suele adaptarse en poco tiempo, gozando después incluso de mayor velocidad de escritura y menor cansancio en sus manos.

puede procesar. Dicha información se convierte en el movimiento de un puntero en pantalla, que refleja el movimiento de la mano.

En primer lugar, el ratón consta de una esfera de material plástico (en adelante, «bola") en su interior, que establece con- tacto con la superficie sobre la que se desliza el ratón (usualmente una alfombrilla diseñada a tal efecto). La bola se puede apreciar en la Figura 1-a. Cuando el usuario desplaza el ratón, la bola

Otros tipos de ratones

El ratón convencional ha evolucionado hacia otras nuevas variantes, con sus ventajas e inconvenientes. En primer lugar, tenemos los ratones inalámbricos. Básicamente, se trata de un ratón convencional, en el cual se ha sustituido el cable de comunicación con el PC por un enlace de radiofrecuencia o infrarrojos. La ventaja radica en que el ratón se puede mover y cambiar de lugar con gran comodidad, ya que no hay un cable que haga difícil dicha tarea. Sin embargo, no hay que olvidar que este tipo de ratón es sensible a señales electromagnéticas. Esto podría conducir a problemas en entornos con un alto nivel de interferencias.

Una opción poco común pero práctica es lo que se denomina *footmouse*. Se trata, en este caso, de un ratón controlado por el pie, en lugar de la mano. La ventaja principal radica en que el teclado se puede usar sin limitaciones (es decir, con ambas manos), mientras se emplea el ratón.

Una tecnología muy utilizada en los ordenadores portátiles es la denominada *Glidepoin*t. Se trata de una pequeña superficie rectangular, donde el usuario desplaza el dedo, y el ratón se mueve de forma acorde en pantalla. Si el usuario desea hacer clic o doble clic, lo puede hacer directamente sobre la superficie, mediante ligeras pulsaciones. Además, se suele disponer de los dos botones típicos del ratón, para los usuarios que desean emplear el método tradicional. Normalmente, los bordes inferior y lateral de la superficie permiten controlar cómodamente las barras de desplaza miento típicas de las aplicaciones para Windows. Como inconveniente, si los dedos del usuario se encuentran húmedos, este tipo de ratón no funcionará correctamente.

Otra tecnología de gran aceptación, lanzada por Microsoft en 1996, es la conocida como *Intellimouse*. Consiste en la introducción de una pequeña rueda, que el usuario puede desplazar en ambos sentidos, y además se puede presionar como un botón más. El desplazamiento de la rueda permite gobernar cómodamente las barras de desplazamiento de Windows. La función asociada a la pulsación de la rueda suele ser programable, asignando tareas como copiar, pegar, etc.

Finalmente, cabe resaltar la existencia de los ratones denominados *trackbal*l. Básicamente, se trata de un ratón convencional, pero en este caso la bola no se mueve como consecuencia del roce con la superficie, sino que es accionada directamente por el usuario. En este caso, ya no es necesario mover el ratón. En la parte superior se encuentra la bola (al alcance del dedo pulgar del usuario) y los botones. En este caso, no es necesario adquirir una alfombrilla y — ya que no hay necesidad de desplazamiento— el ratón requiere poco espacio libre en el área de trabajo. Como se puede intuir, no existe una versión óptica de los ratones *trackball*.

rueda, y hace girar dos pequeños rodillos que se encuentran en contacto con ella (ver Figura 1-b). Uno de los rodillos reacciona al desplazamiento en la dirección X (horizontal), mientras que el otro detecta el desplazamiento en la dirección Y (vertical). Cualquier desplazamiento del ratón se puede entender como la combinación de los desplazamientos horizontal y vertical. Por ello los ejes de giro de los rodillos forman un ángulo de 90 grados.

Cada rodillo se conecta a un eje que hace girar un disco (Figura 1-c). Cada disco presenta perforaciones en su superficie, formando ventanas distribuidas uniformemente. En un lado de cada disco se halla un diodo emisor de infrarrojos (LED de infrarrojos), mientras que en el lado opuesto se encuentra un sensor de infrarrojos (Figura 1-d). Cuando el usuario mueve el ratón, los discos giran. Al desplazarse las perforaciones por delante del LED emisor, se alterna luz y oscuridad en el lado del sensor, es decir, se producen pulsos de luz. El sensor convierte los pulsos de luz en pulsos eléctricos. La señal resultante determina claramente el número de pulsos detectados durante cada periodo de monitorización. Esto permite calcular la velocidad y la longitud del desplazamiento en cada dirección.

Queda una incógnita por resolver: ¿cómo se determina en qué sentido se ha desplazado el ratón en cada dirección? Con la configuración explicada hasta ahora, se puede detectar la distancia recorrida y la velocidad, pero no el sentido del movimiento. Una de las soluciones para resolver este problema es añadir a cada disco un nuevo par emisor-sensor situado justamente en el otro extremo del disco, de tal forma que ambos sensores ven pulsos de luz al mismo tiempo. Entre el disco y cada sensor se coloca una pieza de plástico que presenta una perforación. Dicha pieza actúa como una ventana; en otras palabras, determina lo que cada sensor puede ver. La perforación en uno de los sensores se coloca ligeramente más alta que en el otro sensor. Esto se hace de modo que, cuando un sensor detecta un pulso de luz, el otro está en estado de transición (bien de luz a oscuridad, o viceversa). El proceso (ilustrado en la Figura 2) consiste en centrarse solamente en uno de los sensores (por ejemplo, el sensor A). Cuando se detecta un pulso de luz en A, se observa el tipo de transición que ocurre en el sensor B, pocos instantes después. Si se gira en sentido antihorario, se aprecia que B pasará de luz a oscuridad (pulso negativo). En cambio, si se gira en sentido horario, la señal B pasará de oscuridad a luz (pulso positivo).

Figura 3. La ausencia de partes mecánicas hace que lo ratones ópticos sean más fiables

Visto de otro modo, las dos señales produ-

cidas son iguales, pero aparecen con un cierto retardo de tiempo entre ambas. Según cuál de las dos señales se retarda respecto a la otra, se tiene uno u otro sentido. Este es uno de los métodos para de terminar el sentido, pero no el único.

Un procesador, incluido en el ratón, lee los pulsos y los traduce a información digital, que resulta fácil de procesar por parte del PC. Dicha información se envía al PC en formato serie, a través del cable. No hay que olvidar la presencia de dos o tres botones en el ratón, cuyo estado se incluye en la información enviada al PC.

Ratones ópticos

La firma Agilent Technologies desarrolló en 1999 un tipo de ratón realmente innovador, al que se denominó "ratón óptico". Esta tecnología ha demostrado ser realmente eficaz, y en consecuencia, este tipo de ratón ha gozado (y goza actualmente) de una gran aceptación.

Un ratón óptico (Figura 3) es, básicamente, una pequeña cámara (que toma unas 1.500 imágenes por segundo) y un software de procesamiento digital de imagen en tiempo real.

Se incorpora un diodo emisor de luz (LED) que ilumina la superficie sobre la que se arrastra el ratón. La cámara captura imágenes de la superficie y las envía a un procesador digital de señales (DSP), operando con un rendimiento muy elevado (18 millones de instrucciones por segundo o MIPS). El software que se ejecuta sobre el DSP es capaz de detectar patrones sobre cada imagen recibida. Estudiando cómo se desplazan dichos patrones en las imágenes sucesivas, el DSP averigua el desplazamiento y la velocidad. Esta información se envía al PC cientos de veces por segundo, lo que ofrece una confortable sensación de continuidad para el usuario.

La interfaz entre el ratón y el PC

A nivel de conectores, la mayoría de ratones se comunican con el PC mediante la interfaz PS/2 o conectores para el puerto serie (DB.9, por ejemplo). Independientemente del tipo de conector, el ratón envía al PC tres bytes de información en formato serie, a una velocidad de hasta 1.200 bps. Esto permite enviar información aproximadamente 40 veces por segundo.

El primer byte contiene la siguiente información: estado de los botones izquierdo y derecho, sentido del movimiento en ambas direcciones (X e Y) y la información de desbordamiento en las direcciones X e Y. Los siguientes 2 bytes contienen, respectivamente, el movimiento en las direcciones X e Y. En otras palabras, estos dos bytes contienen el número de pulsos detectados en cada dirección desde la última vez que se envió información al PC. Si el ratón se desliza muy rápido, es posible que se cuenten más de 255 pulsos en cualquiera de las direcciones, y de ahí la inclusión de indicadores de desbordamiento.

Los ratones ópticos reportan varios beneficios en relación con los ratones convencionales. En primer lugar, la ausencia de componentes móviles (bola, discos, etc.) reduce considerablemente la probabilidad de fallos. Tampoco hay

que olvidar que, en los ratones convencionales, la suciedad presente en la superficie de desplazamiento penetra en el interior del ratón con gran facilidad.

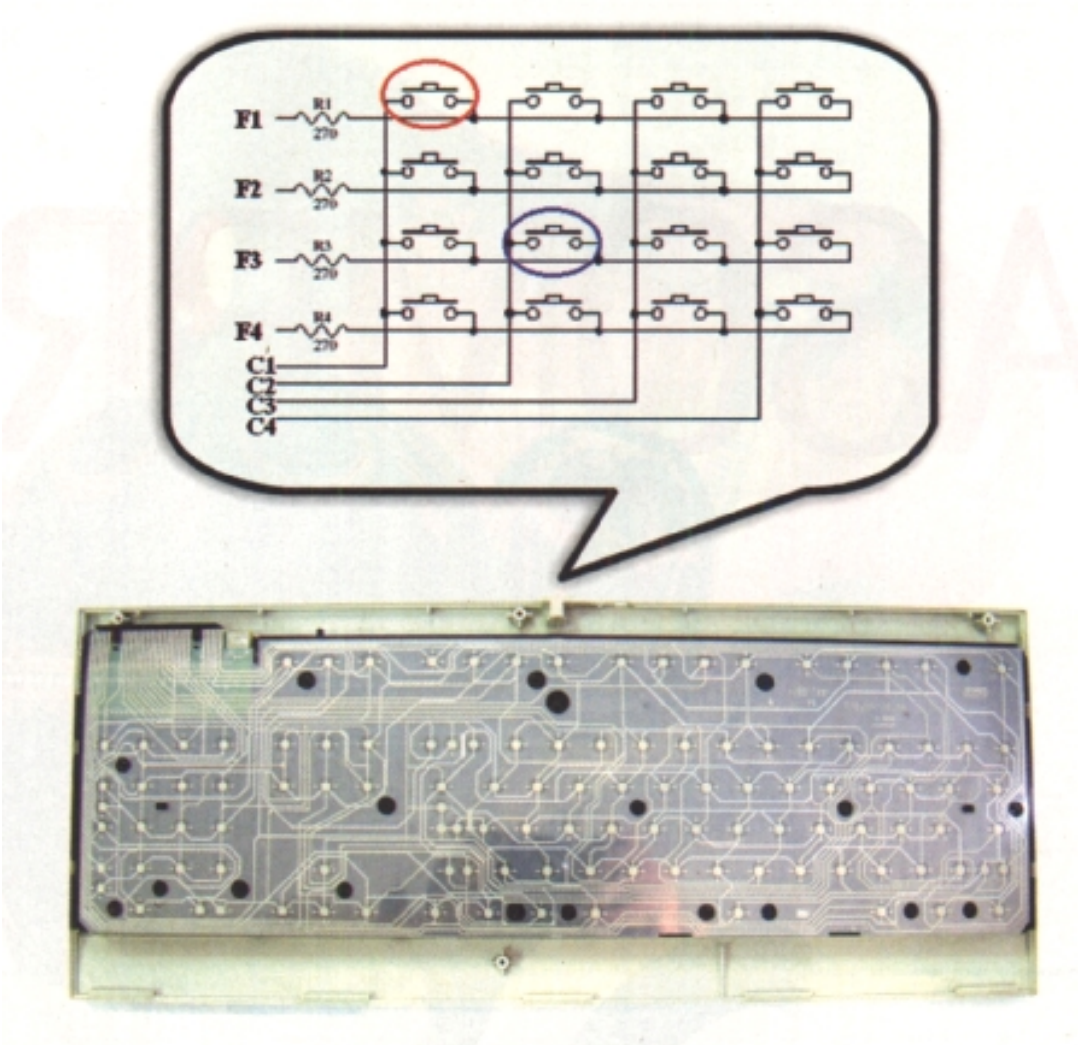


Figura 4. Aspecto y esquema eléctrico de una matriz de teclas.

Esto provoca interferencias en los sensores, algo que no ocurre en los ratones ópticos. Otra ventaja radica en que, en los ratones ópticos, el seguimiento de los movimientos del ratón se realiza a mayor velocidad, obteniendo en pantalla un movimiento más continuo. Finalmente, es importante recalcar que los ratones ópticos no requieren el uso de una superficie especial: en realidad pueden trabajar prácticamente sobre cualquier superficie.

Los primeros ratones ópticos (anteriores a los arriba comentados) se basaban en un LED, que enviaba un haz de luz sobre una superficie especial altamente reflexiva, y un sensor óptico que capturaba el haz reflejado. La superficie presentaba una rejilla de líneas oscuras (que apenas reflejaban luz) sobre ella. Al mover el ratón, el haz de luz era interrumpido un cierto número de veces por las líneas, lo que permitía conocer el desplazamiento y la velocidad.

Este tipo de ratón presentaba diversos problemas de uso. En primer lugar, el usuario debía mantenerlo orientado en un ángulo oportuno, para asegurar un correcto funcionamiento. Además, cualquier daño en la alfombrilla, o la pérdida

de esta, hacía obligatorio adquirir una nueva. Gracias a los nuevos ratones ópticos (comentados arriba), esta tecnología ha sido olvidada y sus problemas han desaparecido.

El teclado

En principio, un teclado no parece presentar demasiados secretos. Aunque no se trata de uno de los componentes más complejos del PC, el teclado es una

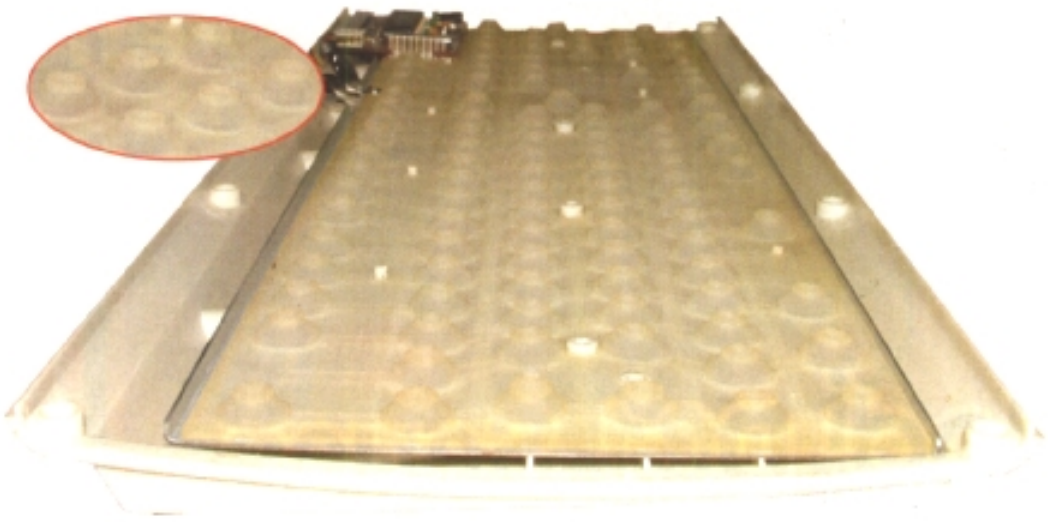


Figura 5. Aspecto de las cúpulas de goma. En la parte superior izquierda se aprecia el circuito controlador.

interesante pieza de tecnología que va un poco más allá de lo aparente. ¿Había usted reparado en que un teclado constituye una pequeña computadora de por sí?

Un teclado es un conjunto de interruptores (teclas), que se hallan conectados a un microprocesador. Este último vigila el estado de los interruptores, y responde de forma específica ante cualquier cambio de estado.

Los teclados suelen incorporar cuatro tipos de teclas: de escritura, de función, de control y de teclado numérico o *keypad*. Las teclas de escritura se suelen organizar en formato QWERTY (son las seis primeras letras que aparecen en este arreglo). La disposición de teclas es justamente la que podemos encontrar en una máquina de escribir.

El teclado numérico (con un total de 17 teclas) facilita enormemente la introducción de dígitos, operadores matemáticos elementales, punto decimal, etc. La disposición es la que podemos encontrar en multitud de calculadoras, lo que hace su uso más familiar.

Las teclas de función, dispuestas en una fila en la parte superior del teclado, permiten que los programas o el sistema operativo les asignen comandos específicos. Por ejemplo, a la tecla F1 se le suele asignar el comando “mostrar ayuda”, casi de forma estándar.

Finalmente, las teclas de control facilitan funciones de edición en pantalla (inicio, fin, insertar, eliminar, escape, etc.) y ofrecen cursores para desplazarse en pantalla. En el caso particular de los teclados diseñados para Windows, aparecen nuevas teclas de control, como “menú inicio” o “menú de contexto”.

Funcionamiento del teclado

El funcionamiento del teclado queda gobernado por el microprocesador y la circuitería de control. Las teclas se hallan ligadas a una matriz de circuitos (o matriz de teclas) de dos dimensiones. Cada tecla, en su estado normal (no presionada) mantiene abierto un determinado circuito. Al presionar una tecla, el circuito asociado se cierra, y por tanto circula una pequeña cantidad de corriente a través de dicho circuito. El microprocesador detecta los circuitos que han sido cerrados, e identifica en qué parte de la matriz se encuentran, mediante la asignación de un par de coordenadas (x,y).

Figura 6. En un Trackball, la bola se mueve directamente con los dedos. En la imagen vemos el Marble Mouse de Logitech.

La Figura 4 muestra el aspecto físico y el esquema de una matriz de teclas. Si se presiona la tecla resaltada en rojo, la corriente fluirá desde F1 hacia C1. El microprocesador identificará la tecla con las coordenadas (1,1), o lo que es lo mismo, fila 1 y columna 1. Si se presiona la tecla resaltada en azul, las coordenadas son (3,2).

Acto seguido, se acude a la memoria ROM del teclado, que almacena lo que se denomina “mapa de caracteres”. Dicho mapa no es más que una tabla que asigna un carácter a cada par (x,y). También se almacena el significado de pulsar varias teclas simultáneamente. Por ejemplo, a la tecla etiquetada como “T” se le asigna el carácter “t”, pero si se pulsa SHIFT simultáneamente, se asigna “T”.

Los teclados permiten que la computadora asigne un nuevo mapa de caracteres, permitiendo crear teclados para multitud de lenguajes.

Como interruptores, las teclas padecen del conocido “efecto rebote”. Cuando una tecla se presiona, se produce una cierta vibración, que equivale a presionar y soltar la tecla repetidas veces, muy rápidamente. Una de las misiones del procesador del teclado es eliminar dicho fenómeno. Cuando el procesador detecta que una tecla cambia de estado con una frecuencia excesiva (mayor que la que un humano puede generar al usar normalmente el teclado), interpreta el conjunto de rebotes como una simple pulsación. Sin embargo, si mantenemos pulsada la tecla más tiempo, el procesador detecta que los rebotes desaparecen, e interpreta que queremos enviar el mismo carácter al PC repetidas veces. La frecuencia con la cual se envía el carácter repetido al PC se puede establecer por software, concretamente desde el sistema operativo.

Tecnologías de teclado

Existen diversos teclados, cuya diferencia se centra en la tecnología empleada para construir los interruptores (teclas). En la actualidad, los teclados más populares emplean teclas de “cúpula de goma”. Las teclas reposan sobre una cúpula fabricada en goma, de pequeño tamaño y gran flexibilidad, con un centro rígido de carbono. Cuando se realiza una pulsación, una pieza colocada bajo la superficie de la tecla hunde la cúpula. Esto hace que el centro de carbono se hunda también, hasta tocar una pieza metálica situada en la matriz de circuitos.

Figura 7. Teclado ergonómico Natural Keyboard de Microsoft.

Mientras la tecla permanezca pulsada, el centro de carbono cerrará el circuito apropiado. Cuando la tecla se libera, la cúpula de goma vuelve a su posición original, y el centro de carbono deja de cerrar el circuito asociado a la tecla. Como consecuencia, la tecla también vuelve a su posición original, quedando lista para volver a ser presionada. Estos teclados resultan económicos y, además, presentan una excelente respuesta táctil. Otra ventaja se centra en su gran resistencia al polvo y la suciedad, ya que las cúpulas de goma aíslan los interruptores. La Figura 5 muestra un teclado de este tipo.

Otro tipo de teclados son los de membrana. Estos se asemejan a los de cúpula de goma en su forma de operar. Sin embargo, en lugar de emplear una cúpula de goma independiente para cada tecla, se basan en una única pieza de goma, que cubre todo el teclado y contiene un abombamiento para cada tecla. Estos teclados no se encuentran con facilidad en el mundo de los ordenadores personales, ya que ofrecen una respuesta táctil inapropiada. En cambio, gracias al

gran aislamiento al que se somete la matriz de circuitos, estos teclados se emplean habitualmente en sistemas sometidos a condiciones extremas. Pasando a una tecnología no mecánica, encontramos los teclados capacitivos. En estos, los interruptores no son realmente mecánicos: de hecho, la corriente fluye continuamente por toda la matriz de teclas. Cada tecla está provista de un muelle, que asegura el retorno a su posición original tras una pulsación. Bajo la superficie de cada tecla se halla una pequeña placa metálica. Bajo dicha placa, a una cierta distancia, se halla otra nueva placa metálica. El conjunto de dos placas metálicas separadas por un material dieléctrico (el aire, en este caso) no es más que un condensador. La capacidad de dicho condensador varía en función de la distancia entre las placas. Por tanto, al pulsar la tecla (y por tanto acercar las placas), se produce un cambio de capacidad que sirve para detectar la pulsación de la tecla. El coste de estos teclados es elevado pero, por otro lado, se deterioran muy poco. Esto último les permite gozar de una larga vida, mayor que la ofrecida por cualquier otra tecnología de teclados. Ya que las dos placas nunca entran en contacto directo, no existen rebotes, lo que supone otra ventaja importante.

Otra tecnología más simple es la de contacto metálico. En ella, las teclas se dotan de un resorte, y cada circuito se cierra por el contacto directo entre dos placas metálicas. Otra variante introduce un material esponjoso entre las dos placas. En general, esta tecnología proporciona una buena respuesta táctil. El problema reside en que los contactos se deterioran rápidamente, ya que no existe una barrera aislante que proteja la matriz de contactos, como en los teclados de membrana o cúpula de goma