

ESCUELA UNIVERSITARIA DE INFORMÁTICA DE SISTEMAS
UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

Apuntes de la asignatura de:

PERIFÉRICOS

Preparados por:

Juan Carlos Lázaro Obensa

Dpto. de Informática y Automática

INDICE

1.- INTRODUCCIÓN	1
1.1 ¿Que son los periféricos?	1
1.2 Aspecto comercial de los periféricos	5
1.3 Enlace con el procesador	5
1.4 ¿Cómo ve el software a los periféricos?	6
1.5 Visto por el operador	8
1.6 Periféricos de E/S y periféricos de almacenamiento	8
 2.- PERIFÉRICOS DE ALMACENAMIENTO	 11
2.1 Introducción	11
2.2 Tambores y discos magnéticos	12
2.3 Cabezas por pista y movimiento de las cabezas	13
2.4 Tambores y discos	13
2.5 Discos y cabezas múltiples	14
2.6 Espaciado entre cabezal y disco	16
2.7 Diseño de cabezas	18
2.8 Posicionamiento de la cabeza	21
2.9 El medio magnético	22
2.10 Grabación de pulsos. Precompensación	24
2.10.1 Superposición lineal. Precompensación	25
2.11 Optimización del espacio. "Banding"	26
2.11.1 Múltiples bandas	28
2.12 Formato de grabación	30
2.13 Servopistas	35
2.14 Formato de la pista	36

2.15	Entrelazado	38
2.16	Tratamiento de errores	39
2.17	Cálculo del CRC	42
2.17.1	División polinómica por hardware	42
2.17.2	Aritmética en módulo 2	43
2.17.3	División larga en módulo 2	43
2.18	Formato de alto nivel	44
2.19	Organización del disco en el S.O. DOS	45
2.19.1	Estructura lógica del disco	45
2.19.2	Organización de los discos	46
2.19.3	El registro de arranque (BOOT)	46
2.19.4	Tabla de localización de ficheros	47
2.19.5	El directorio	48
2.19.6	El espacio de datos	50
2.20	El almacenamiento óptico	50
2.20.1	El sistema óptico	51
2.20.2	Seguimiento de la pista	54
2.20.3	Control de enfoque	56
2.20.4	Rotación del disco	58
2.20.5	Formatos de grabación	59
2.21	Un nuevo formato: el DVD	61
3.-	INTERFACES SERIE Y PARALELO.....	63
3.1	Introducción	63
3.2	Problemas en las transmisiones serie	64
3.2.1	Sincronización de bit	64
3.2.2	Sincronización de carácter	66
3.2.3	Sincronización de mensaje	66
3.3	Métodos de E/S para comunicaciones serie	67
3.3.1	Método asíncrono	67
3.3.2	Método síncrono	68
3.3.3	Regeneración del reloj en el receptor	68
3.4	Estándar de comunicación serie RS-232	70
3.4.1	Variantes RS-422, 423 y 485	74
3.5	El interfaz MIDI	76
3.5.1	Un poco de historia	76
3.5.2	El hardware MIDI	77
3.5.3	Protocolo de mensajes de MIDI	79
3.6	Interfaces paralelo	81
3.7	El interfaz ST-506/412	81
3.7.1	Generalidades	81
3.7.2	Cableado	82
3.7.3	Señales y funcionalidad	82
3.7.4	Ejemplo de implementación: La tarjeta controladora WD1003-WAH	84

3.8 Interfaz ESDI	86
3.9 Bus SCSI.....	87
3.9.1 Generalidades.....	87
3.9.2 Señales y funcionalidad.....	88
3.9.3 Fases del bus SCSI	90
3.9.4 Fases de transferencia de información.....	92
3.9.5 Variantes síncrona y ancha	94
3.9.6 Condiciones especiales del bus	94
3.10 Los interfaces Centronics e IEEE-1284	95
3.10.1 Introducción y necesidad de la norma	95
3.10.2 Modo compatible (Centronics convencional).....	97
3.10.3 Modo nibble	98
3.10.4 Modo byte	99
3.10.5 Modo EPP (Enhanced Parallel Port)	100
3.10.6 Modo ECP (Extended Capability Port)	102
3.10.7 Negociación de modo.....	104
3.11 Bus IEEE-488	106
3.11.1 Estructura del bus	107
3.11.2 Examen funcional del bus.....	108
3.11.3 Protocolo de operación.....	111
4.- PERIFÉRICOS DE ENTRADA.....	115
4.1 Teclados	115
4.2 Tipos de pulsadores	116
4.2.1 Pulsador de lámina flexible	118
4.2.2 Pulsador de bovedilla	118
4.2.3 Pulsador elastómero	119
4.2.4 Pulsadores Reed	119
4.2.5 Pulsadores capacitivos.....	120
4.2.6 Pulsador de efecto Hall	120
4.2.7 Pulsador inductivo.....	122
4.3 Codificación.....	122
4.3.1 Conexión a codificador	123
4.3.2 Conexión matricial	124
4.3.3 Exploración secuencial.....	126
4.3.4 Codificación por microprocesador	127
4.3.5 Doble codificación	128
4.4 Pulsación simultánea de varias teclas	128
4.4.1 Sobrepulsación de dos teclas.....	129
4.4.2 Inhibición de N teclas.....	129
4.4.3 Sobrepulsación de N teclas.....	129
4.5 Ratones y tabletas gráficas.....	129
4.5.1 Funcionamiento básico del ratón.....	130
4.5.3 Tablet gráficas.....	132
4.6 Lectores de código de barras.....	132
4.6.1 Simbología de códigos de barras.....	133
4.6.2 Equipamiento de lectura.....	135

5.- SISTEMAS DE VÍDEO.....	139
5.1 Introducción	139
5.2 Generación de la imagen en un TRC	140
5.3 Estudio de un visualizador de barrido secuencial	142
5.3.1 La pantalla del visualizador.....	142
5.3.2 Sincronismo horizontal o señal H.....	145
5.3.3 Sincronismo vertical o señal V.....	145
5.3.4 Señal de modulación de la intensidad del haz o señal Z.....	145
5.3.5 Magnitudes significativas	147
5.4 Tipos de monitores	148
5.4.1 Monitores mono y multi-frecuencia	148
5.4.2 Monitores analógicos y digitales	149
5.4.3 Entrelazado.....	149
5.4.4 Monitores de color	149
5.5 Controlador de pantalla.....	150
5.5.1 Memoria de pantalla.....	151
5.5.2 El procesador gráfico	152
5.6 Generación de la señal de video.....	153
5.6.1 Generador de caracteres	153
5.6 Ejemplos de tarjetas	157
6.- PERIFÉRICOS DE SALIDA	175
6.1 Introducción	175
6.2 Impresoras de impacto	176
6.2.1 Máquinas de escribir y teletipos	176
6.2.2 Impresoras de margarita	178
6.2.3 Impresoras de barril.....	180
6.2.4 Impresoras de banda de cadena y de tren	181
6.2.5 Impresoras de matriz de puntos	182
6.2.6 Impresoras de matriz de líneas	184
6.2.7 Impresoras de color de matriz	185
6.3 Impresoras de NO impacto	186
6.3.1 Impresoras de chispa electrostática	186
6.3.2 Impresoras electroquímicas	186
6.3.3 Impresoras térmicas.....	187
6.3.4 Impresoras electrográficas.....	187
6.3.5 Impresoras Láser	188
6.3.6 Impresoras LED, LCD y de deposición de iones.....	189
6.3.7 Impresoras magnetográficas	190
6.3.8 Impresoras de inyección de tinta	190
6.3.9 Plotters de plumas	190
6.3.10 Plotters electrostáticos.....	192
6.4 Dithering o entramado	193

7.- SISTEMAS DE INSTRUMENTACIÓN Y CONTROL.....	195
7.1 Transductores y señales de campo	195
7.1.1 Transductores de resistencia variable	196
7.1.2 Transductores de reactancia variable (capacitivos o inductivos).....	196
7.1.3 Transductores generadores de carga.....	197
7.1.4 Transductores generadores de tensión.....	197
7.1.5 Transductores generadores de corriente	197
7.1.6 Transductores digitales.....	197
7.2 Sistemas de adquisición de datos.....	197
7.2.1 Introducción	197
7.2.2 Cuantificación y codificación.....	199
7.2.3 Muestreo y "aliasing"	200
7.3 Circuitos básicos de un sistema de adquisición de datos.....	202
7.3.1 Amplificadores	202
7.3.2 Codificación digital	203
7.3.3 Conversores digitales/analógicos (D/A)	204
7.3.4 Conversores analógico-digitales.....	206
7.3.5 Multiplexores analógicos	209
7.3.6 Circuitos de muestreo y retención	210
7.3.7 Modos de conexión de un sistema de adquisición de datos a un ordenador	211
7.3.8 Especificaciones y parámetros característicos	212

Introducción

1.1 ¿QUE SON LOS PERIFÉRICOS?

El diccionario de la computación de Oxford define a los periféricos como 'algún dispositivo, incluyendo dispositivos I/O y almacenamiento que son conectados al ordenador'. Otra posible definición es 'máquinas que pueden ser operadas bajo el control de una computadora'.

'Periférico' es la contracción del término 'Dispositivo periférico', el cual podríamos utilizar siempre si no fuésemos perezosos, ya que dispositivo periférico es por supuesto un dispositivo que está en la periferia del ordenador, distinto del procesador central y la memoria principal, los cuales constituyen la unidad básica del ordenador. Los dispositivos periféricos son en alguna medida opcionales de tal forma que un periférico específico no es esencial para el ordenador, pero éste vería restringido su utilidad si no tuviera todos los periféricos. De esta forma, los periféricos son bastante fáciles de reconocer cuando vamos a comprar un ordenador. Son los 'extras' que debemos añadir al ordenador básico. Algunos dispositivos periféricos pueden estar incluidos en la misma caja que el ordenador básico e incluidos en su precio. Diferentes modelos pueden variar sólo según la selección de los dispositivos periféricos que incluyan.

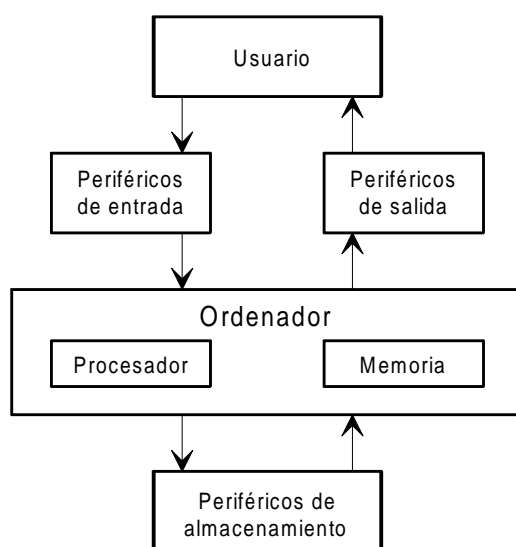


Fig. 1.1 Periféricos de entrada, salida y almacenamiento

Desde el punto de vista funcional podemos dividir los periféricos en dos importantes clases, según se muestra en la figura 1.1, en la que podemos diferenciar los dispositivos que comunican al ordenador con el exterior (dispositivos de Entrada/Salida) y los que almacenan los datos fuera de la memoria principal del sistema. El término 'datos' es usado aquí en su sentido más general, incluyendo también programas. En algunos de estos dispositivos los datos pueden ser borrados (por ejemplo los discos flexibles), mientras que en otros los datos son una parte permanente del dispositivo (como por ejemplo los discos ópticos). Hay también casos en los que los dispositivos de almacenamiento son intercambiables y pueden ser considerados también como dispositivos de entrada/salida al permitir introducir en el sistema datos provenientes de otra máquina y almacenar datos que puedan ser introducidos en otro equipo. Este tipo de dispositivos, también está más cerca del usuario, ya que es éste el que habitualmente realiza los intercambios. Sin embargo, no pueden ser considerados como dispositivos de interfaz con el usuario ya que la información almacenada no es directamente interpretable por el usuario y requiere siempre una máquina para visualizarla, al contrario de lo que sucede con las pantallas o las impresoras.

La frontera entre los dispositivos periféricos y el ordenador central no está totalmente definida en muchos casos ya que debe existir la interfaz necesaria, compuesta a veces por dispositivos más complejos tales como canales, controladores de bus, procesadores periféricos específicos etc. Según esto, un determinado componente lo consideramos o no periférico según la parte del sistema que consideremos como unidad central. En su condición más extrema, periférico sería todo aquel dispositivo o componente externo a la CPU y que no es necesario para el funcionamiento del ordenador como tal. Sin embargo, determinados dispositivos, están tan íntimamente relacionados con la arquitectura y comportamiento del sistema que su funcionamiento y especialmente su rendimiento se vería considerablemente mermado sin ellos. A este grupo pueden pertenecer los elementos que se encargan del control de las interrupciones, controladores de DMA, controladores de bus, etc. sin los cuales no es posible concebir un ordenador actual. No obstante no dejan de ser elementos externos a la CPU, opcionales (al menos en la fase de diseño del sistema) y encargados cada uno de ellos de una tarea específica y siempre bajo el control del procesador central. Esta consideración de externos a la CPU debe entenderse desde una perspectiva funcional, ya que físicamente pueden estar incluidos en el mismo circuito integrado del microprocesador, precisamente porque como se ha comentado, su influencia en el comportamiento global del sistema es tan decisiva, que resulta difícil imaginar un ordenador actual, por simple que sea, que no disponga de tales elementos. En el extremo opuesto, desde el punto de vista constructivo, podríamos considerar como periférico, a todo aquel dispositivo opcional externo a la carcasa que contiene la unidad central de proceso, como puede ser un monitor de video o una impresora. Sin embargo, esta definición tiene sus problemas. Imaginemos el caso de una unidad de almacenamiento masivo, tal como una unidad de cinta, un dispositivo magneto-óptico, o un disco duro; cualquiera de estos dispositivos puede estar incluido en la carcasa de la CPU y también existen versiones externas de los mismos. Obviamente tanto la versión externa como la interna de estos dispositivos es esencialmente la misma, salvo consideraciones de cables de conexión y fuente de alimentación independientes, con lo que resulta chocante que según elijamos una u otra versión el dispositivo es o no periférico. No obstante esta definición podría ser válida para un usuario elemental.

Queda claro pues que la definición de periférico viene condicionada por el nivel en el que vayamos a trabajar. Lo que para un diseñador de sistemas puede ser un periférico (por ejemplo el controlador de interrupciones), un usuario ni siquiera sabe de su existencia. En un camino intermedio se situarían los usuarios avanzados o programadores, que distinguirán niveles intermedios de periferia.

A lo largo de este texto vamos a considerar una opción intermedia global, centrándonos en el concepto de periférico, como aquel dispositivo que permite una comunicación de la unidad central de proceso con el exterior. Respecto a esta definición maticemos dos conceptos clave que aparecen en la misma. Por una parte, el concepto de comunicación debe entenderse en su sentido

más general y no solo en el ámbito de las comunicaciones telemáticas, en las que la información se intercambia entre varios sistemas computadores. También deben incluirse por tanto, aquellos dispositivos que permitan una comunicación con el usuario, tales como monitores de video, impresoras, etc. Y respecto al concepto de exterior, consideraremos como tal todo aquello que esté más allá de la circuitería específica de procesamiento y funcionamiento básico, independientemente de que esté ubicado dentro o fuera de la carcasa de la CPU. Es decir, todo aquello que esté más allá del bus principal.

Una definición más precisa podría establecerse teniendo en cuenta que los procesadores disponen de instrucciones y/o señales de control del bus que son específicas para realizar operaciones de entrada/salida. Desde este punto de vista un dispositivo periférico será todo aquel que requiera este tipo de instrucciones o de señales de control para intercambiar información con la CPU (Unidad Central de Proceso).

Una vez vistas todas las definiciones o aclaraciones anteriores, podemos continuar ahora con la siguiente pregunta, ¿por qué necesitan los ordenadores a los periféricos?. Claramente, los periféricos de Entrada /Salida (E/S) son para enlazar al ordenador con el medio exterior, con el usuario humano, y más allá del objetivo de este texto para la comunicación entre máquinas. Sin los dispositivos periféricos E/S, no habría caminos para instruir al ordenador (asignarle tareas), no podríamos comunicarle datos y tampoco sería posible conocer los resultados por él producidos. De esta forma, los periféricos de E/S (y particularmente teclados, pantallas e impresoras) son de fundamental importancia en los sistemas computadores.

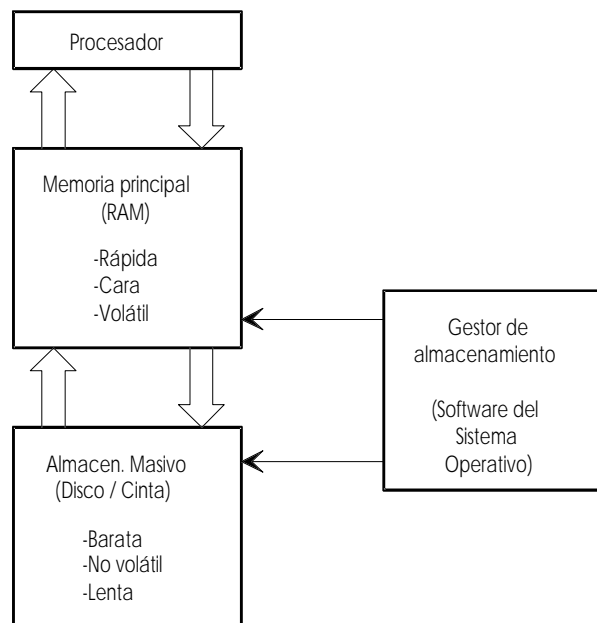


Fig. 1.2 Almacenamiento en dos niveles

La necesidad de los dispositivos de almacenamiento, tanto de tipo masivo convencional como de respaldo, resulta obvio. El almacenamiento principal del computador, conocido también como memoria RAM (Memoria de Acceso Aleatorio), tiene que ser de acceso rápido y permitir accesos aleatorios a su contenido siendo este tipo de almacenamiento caro. Para tener unas prestaciones razonables, algunos o todos los programas y datos que están en uso en algún momento deben ser retenidos en el almacenamiento principal; pero la mayoría de los usuarios tienen todavía más programas y datos que no son usados en ese instante, pero que deben ser almacenados dentro del sistema para que estén disponibles en poco tiempo. Esto nos hace limitar el tamaño de la memoria principal, que es cara, y almacenar estos datos inactivos en sistemas menos caros y que denominamos medios de almacenamiento masivo (fig. 1.2). Datos y programas son movidos (o

copiados) entre estos y la memoria principal cuando son necesarios. La mayoría de las veces cuando un programa ha concluido su tarea otro ocupa el espacio que ocupaba en memoria principal.

Una razón para el uso de medios de almacenamiento masivo es que el almacenamiento principal es volátil, lo que quiere decir que todo lo almacenado en él se pierde cuando la alimentación es cortada o falla. Los datos son por lo tanto copiados o grabados en los medios de almacenamiento masivo no volátiles, tan pronto como los programas terminan con ellos. Los programas no necesitan ser grabados, puesto que normalmente no serán modificados, de tal forma que si ellos fuesen copiados en la memoria principal desde el medio de almacenamiento masivo, ahí permanecerá el original, desde donde se podrá recuperar tantas veces como sea necesario. Este concepto de dos niveles de almacenamiento fue introducido bastante pronto en el desarrollo de las computadoras de propósito general, y es casi universal hoy día. Ocasionalmente hay más de dos niveles. Algunas veces, el usuario es consciente directamente o indirectamente de los dos niveles; por ejemplo, comenzando un programa tecleando su nombre por el teclado hace que ese programa sea transferido desde el medio de almacenamiento masivo a la memoria principal antes de ser ejecutado. Análogamente los comandos de salvar o guardar hacen que los datos sean copiados desde la memoria principal al medio de almacenamiento masivo para un uso posterior. Sin embargo, la mayoría de las computadoras modernas van de alguna forma hacia la idea de almacenamiento virtual donde el usuario no es consciente de la estructura de dos niveles. Para el usuario aparece entonces sólo un nivel (almacenamiento o memoria virtual), y todas las transferencias entre niveles son realizadas automáticamente por el sistema operativo (el cómo lo hace no es objetivo de este curso).

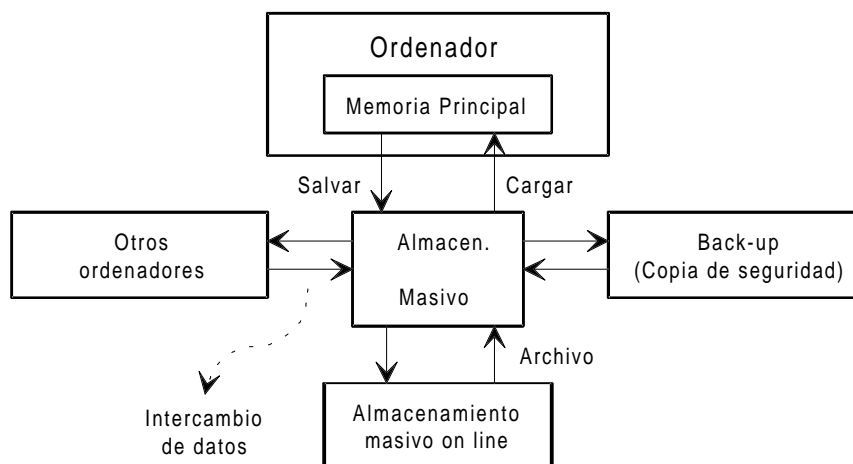


Fig. 1.3 Jerarquía de almacenamiento

Los periféricos de almacenamiento no se usan solamente para almacenamiento masivo sino que pueden ser usados como un sistema intermedio en el movimiento de software y datos desde una máquina a otra según se indica en la figura (1.3). En estas aplicaciones se han comenzado a utilizar dispositivos de E/S, aunque es conveniente tratarlos aún como periféricos de almacenamiento. De hecho, quizás el mismo dispositivo sirve como medio de almacenamamiento masivo; los disquetes en los ordenadores personales (PC) son un obvio ejemplo. Otros dispositivos de almacenamiento pueden ser usados para copias de seguridad o de respaldo ('backup') del almacenamiento masivo, realizando un duplicado de los datos existentes por si los datos del medio de almacenamiento masivo se pierden debido a un fallo del sistema o por un error del programa o del operador. Es importante señalar que no debemos confundirnos entre los términos almacenamiento masivo y de respaldo. El almacenamiento masivo es una parte del trabajo de almacenamiento del ordenador; los datos son grabados cuando ellos son copiados desde la memoria principal al medio de almacenamiento masivo y aquí permanecen hasta que sean utilizados de nuevo. En contraste, los datos son copiados desde el medio de almacenamiento

masivo hacia un nivel superior o memoria de respaldo y no es necesario sustituirlos a menos que exista un problema. Esto supondría un tercer nivel en la jerarquía de almacenamiento y normalmente requiere la intervención del operador, ya que las copias de seguridad se almacenan fuera del sistema, en algunos casos, en armarios blindados e ignífugos para que resistan cualquier eventualidad.

Un uso más de los periféricos de almacenamiento son los almacenamientos fuera de línea ('off-line'), cuando los medios donde se almacenan los datos (tales como discos flexibles) pueden ser separados del dispositivo periférico y almacenados lejos del dispositivo que los grabó. Esto reduce considerablemente el presupuesto de almacenamiento, de tal forma que sólo los medios y no los mecanismos soportes son dedicados a un tipo de datos en particular. Un caso particular es cuando los datos se vuelvan viejos y raramente sean usados, pero no obstante deben ser guardados para el caso en que sean requeridos de nuevo. Es la solución que se adopta habitualmente para las copias de seguridad.

1.2 ASPECTO COMERCIAL DE LOS PERIFÉRICOS

Escoger un ordenador y el software que lo soporta puede llevarnos muchas horas y esfuerzo. El escoger el periférico se considera a menudo como una tarea menor. Eso se vuelve una sorpresa para algunos usuarios que observan que los periféricos tienen un coste por encima de la mitad del coste del hardware total del ordenador. Como ejemplo tomemos un pequeño ordenador, la mayoría de los fabricantes dicen que su unidad básica es mucho mejor que la de su competidor, pero el usuario también necesitará una pantalla, un teclado, y en menor medida discos y usualmente impresora; todo esto junto costará más de la mitad del sistema unitario, y tendrá un mayor efecto en el rendimiento del sistema en la mayoría de las aplicaciones. De esta forma, algunos usuarios (y diseñadores de sistemas) conocen mucho menos de los periféricos que de su procesador central y memoria principal.

1.3 ENLACE CON EL PROCESADOR

El límite preciso entre dispositivos periféricos y la unidad básica no es siempre fácil de reconocer. Una circuitería física, no un software, es necesario para convertir las instrucciones y datos manejados por el procesador central del computador en una forma en la que pueda ser usada por el periférico. Esta circuitería lógica puede ser dividida a menudo en una sección que es común a muchos tipos de periféricos y otra sección que es específica de un tipo de periférico. Una o ambas de estas secciones pueden estar contenidas en el 'sistema central', y en realidad la primera sección puede estar completamente integrada con ella. La segunda sección está a menudo contenida en el dispositivo periférico, o puede estar en una unidad separada.

Un periférico puede requerir cuatro tipos de recursos de la Unidad Central: Memoria, direcciones de I/O, líneas de interrupción y algún canal de DMA (Acceso Directo a Memoria). Un periférico particular puede no necesitarlos todos, dependiendo de sus características y prestaciones. Algunos de estos recursos son abundantes y pueden ser asignados sin muchos problemas a un periférico particular, sin embargo otros son escasos y requieren una asignación más elaborada ya que por regla general, un mismo recurso no puede ser asignado a más de un dispositivo, ya que podrían producirse conflictos que ocasionarían un mal funcionamiento de todo el sistema.

Estas asignaciones de recursos pueden resultar un quebradero de cabeza cuando se añade un nuevo dispositivo al sistema. Esto ha provocado que durante los últimos años se hayan dedicado

considerables esfuerzos por parte de los fabricantes para minimizar los conflictos que aparecen al realizar el reparto de recursos entre distintos periféricos. Como resultado, han surgido las especificaciones Plug & Play (Enchufar y Listo) para ordenadores personales tipo PC y especialmente el bus PCI que lo incorporan actualmente diferentes sistemas con arquitecturas distintas.

Son los fabricantes de ordenadores los que permite escoger los periféricos que se pueden conectar a cada computadora y sus características. El usuario puede también cambiar los periféricos de una computadora a otra, lo cual significa reemplazar el procesador, corazón del sistema, sin tener que comprar un juego completamente nuevo de periféricos. Para hacer esto fácil, los constructores definen interfaces entre la unidad básica y sus periféricos. Cada uno de los nuevos procesadores centrales, y cada uno de los nuevos periféricos pueden ser entonces diseñados para emparejarse a una interfaz predefinida. Cuando un fabricante quiere que todos sus clientes compren sus periféricos, utiliza comúnmente interfaces que son exclusivos de esta compañía y no publica sus especificaciones. Afortunadamente, ésta actitud se está volviendo menos común, y la mayoría de los ordenadores usan interfaces que están publicadas y a menudo desarrolladas por un consorcio o conjunto de empresas que se comprometen a apoyar la norma fabricando equipos que la utilicen. Este tipo de interfaces son publicados como normas estándar, aprobadas por distintos organismos de normalización como IEEE, ANSI, ISO o CCITT quedando por tanto controlados por conjuntos independientes y disponibles para todos los fabricantes que lo deseen. Así, el usuario puede comprar su computadora y periféricos de fabricantes diferentes.

La interfaz de periféricos debe situarse en alguna parte entre el bus principal de la unidad básica y el mecanismo del dispositivo periférico, pero dentro de estos límites hay un amplio margen de variación. Puede haber igualmente dos interfaces en cascada; de esta forma, la circuitería lógica que enlaza el dispositivo periférico a la unidad básica es escindida en tres, no en dos secciones. En este caso, la sección cercana a la unidad básica serviría para todo tipo de periféricos; la sección intermedia que puede describirse como 'interfaz adaptadora', puede concebirse para un tipo de periféricos -por ejemplo discos magnéticos- y la tercera estará dedicada a un modelo particular de ese tipo.

Esa parte del circuito lógico, que controla a un dispositivo periférico es usualmente considerada como una parte de ese dispositivo y se describe como dispositivo controlador para distinguirla del mecanismo del dispositivo (partes electromecánicas). Sin embargo, la división no siempre es tan clara como ésta. Aunque el dispositivo controlador está a menudo dentro de la misma carcasa que el mecanismo, no es necesariamente así siempre, particularmente cuando el controlador es compartido por dos o más mecanismos. Incidentalmente, el término controlador periférico es ambiguo y por lo tanto es mejor evitarlo. En algunos casos se utiliza para describir el dispositivo controlador y en otros para describir la unidad que se encuentra más cerca de la unidad básica y soporta la mayoría de sus periféricos.

1.4 ¿CÓMO VE EL SOFTWARE A LOS PERIFÉRICOS?

En las primeras computadoras las ventajas de los interfaces y protocolos definidos entre la unidad básica y los periféricos no fueron siempre suficientemente apreciados, y en algunos casos el programador de aplicaciones tuvo que conocerlo todo acerca del modo de trabajo de cada periférico. Por ejemplo, había que estar seguro del tiempo que ha de transcurrir entre sucesivos comandos del dispositivo lo que constituía un inconveniente. Una solución que es más o menos universal hoy día, fue delegar el manejo de todos los periféricos al sistema operativo. Los detalles varían de un sistema operativo a otro, pero en general la aplicación llamará al sistema operativo dándole el dato a transferir entre la memoria principal y el periférico especificado. Cuando hay varios periféricos del mismo tipo, el sistema operativo puede decidir cual de ellos es usado.

Una tarea del sistema operativo es identificar un dato en términos de su dirección en la memoria principal. También puede necesitarse determinar la dirección del dato dentro del dispositivo periférico (por ejemplo en discos magnéticos) aunque esto no es necesario con dispositivos serie, tales como impresoras. La dirección puede ser la localización actual en el almacenamiento medio: dirección física. Sin embargo, esto significa que el sistema operativo debe conocerlo todo acerca del dato en el medio de almacenamiento (el cual varía desde un dispositivo a otro), y deben tenerse en cuenta también posibles defectos en el medio de almacenamiento debido al movimiento de datos y algunos otros de localización. Las últimas tendencias se han dirigido hacia dispositivos periféricos inteligentes, particularmente los medios de almacenamiento masivo. Éstos detectan sus propios defectos y almacenan un registro detallado de la localización de los mismos dentro del dispositivo. El sistema operativo sólo necesita conocer la dirección lógica. El sistema operativo ve al periférico como un dispositivo mucho más simple, consistiendo en una secuencia simple de bloques lógicos de longitud fija. La gran ventaja de esto es que el sistema operativo se vuelve más independiente del diseño de periféricos. Todo lo necesario para conocer el número y tamaño de los bloques lógicos y funciones -tales como lectura y escritura- puede realizarlo el dispositivo. Los periféricos inteligentes permiten que el sistema operativo les interroge para conseguir información, ya que ésta no necesita ser construida dentro del sistema operativo.

Tanto si el periférico es inteligente o no, el sistema operativo no necesita conocer la forma física de la interfaz entre la unidad básica y los periféricos; como puede ser el cableado usado, niveles de voltajes que representan las señales, etc. ya que todo eso está dentro del hardware. Ese hardware puede incorporarse en realidad al microprocesador con su propio software o 'firmware' (microprogramas almacenados permanentemente) pero esto no nos concierne ahora. Hasta ahora, tal y como hemos concebido el sistema operativo, los dispositivos periféricos aparecen como una serie de registros y una posible fuente de interrupciones; el sistema operativo pone comandos y direcciones en estos registros, o en algunos de ellos y el periférico vía interfaz hardware, da la respuesta (a menudo llamado 'status') en el mismo registro, o algunos diferentes, dándole al sistema operativo la información que pueda necesitar.

El sistema operativo no es afectado normalmente por la transferencia de datos entre periférico y memoria ya que eso es manejado automáticamente por el dispositivo y el soporte hardware. En muchos sistemas este hardware tiene un camino especial conocido como 'Direct Memory Access' o DMA separado del usado por el procesador central según se indica en la figura 1.4.

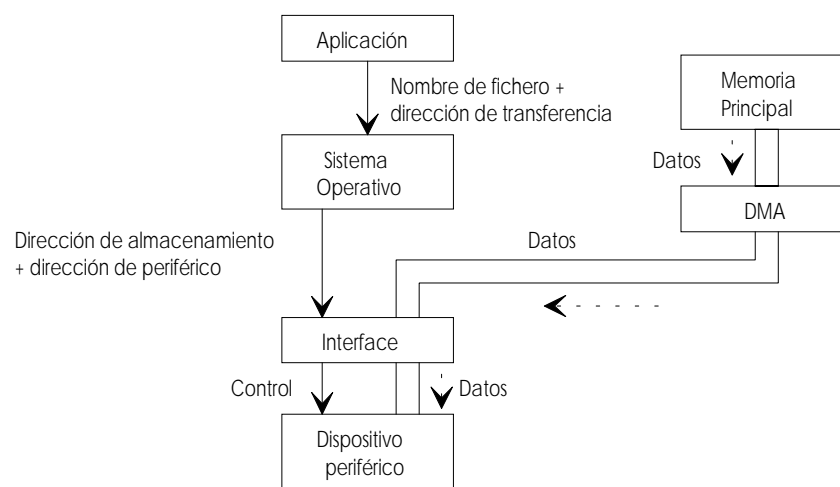


Fig. 1.4 Transferencia de datos al periférico de almacenamiento

Sin embargo, el sistema operativo (y algunas veces las aplicaciones software) necesitan conocer cuando se han completado las transferencias. El sistema puede especificar esto teniendo el tiempo de transferencia en el 'firmware' de la máquina o del dispositivo (set-up de arranque). Sin embargo resulta mucho más eficiente la generación de una interrupción para notificar la finalización de una transferencia, pero alternativamente un estado puede ser activado; esto es un bit simple en el registro o localización de memoria. El sistema operativo puede escoger un instante adecuado para mirar el estado, normalmente justo antes de cada nuevo acceso.

El número, tamaño y nombre de los registros varía desde un sistema a otro. y pueden encontrarse dos soluciones distintas que son conocidas como mapeados en memoria y mapeada en puertos o mapeado independiente. En la primera solución, el sistema dispone de un único espacio de direccionamiento compartido para la memoria y para los puertos de entrada/salida; cualquier dirección puede ser utilizada para memoria o para un registro de E/S. Los sistemas que incorporan la segunda solución, disponen de un espacio de direccionamiento para memoria y de otro independiente para los registros manteniéndose separados a nivel físico. El acceso a uno u otro espacio de direccionamiento va gestionado por las señales de control. Tanto en los sistemas que utilizan puertos mapeados en memoria como aquellos que emplean mapeado independiente disponen de instrucciones distintas para acceder a puertos o a memoria, lo que permite activar distintas señales de control en el bus.

1.5 VISTO POR EL OPERADOR

Los programas de aplicación pueden no necesitar conocer qué dispositivo periférico está enlazado con él, pero el usuario -tanto si es un operador a tiempo completo como si es un usuario casual- tiene una interacción más próxima por lo que es conveniente que el usuario se comunique con el ordenador de forma conveniente. Los dispositivos de E/S existen principalmente para comunicarse con los humanos. Es igualmente importante para el usuario conocer qué se puede esperar de ellos. Los periféricos de almacenamiento son menos interesantes para el operador, especialmente si el medio de almacenamiento no es intercambiable. Por ejemplo, el usuario de un computador personal puede ignorar la presencia de un disco duro fijo, excepto quizás si se ve la luz intermitente, lo cual significa muy poco para él. Donde el medio de almacenamiento puede ser extraído, como los discos flexibles, las cintas magnéticas o los discos magneto-ópticos, el operador es consciente de la existencia dispositivo. Como en comentarios anteriores, en este caso el dispositivo de almacenamiento está en efecto comenzando a ser utilizado como dispositivo de E/S.

Al comienzo, cuando todas las computadoras tenían operadores cualificados, los periféricos no siempre eran fáciles de usar. Además de la dificultad para cargar el papel en las impresoras y de las cintas magnéticas, el operador tenía un número de control para cada dispositivo y tenía que conocer la secuencia correcta para su uso. Hoy día, la mayoría de los periféricos han sido contruidos para ser usados fácilmente, particularmente por el uso de cartuchos facilmente intercambiables. También los controles manuales han sido simplificados o eliminados.

1.6 PERIFÉRICOS DE E/S Y PERIFÉRICOS DE ALMACENAMIENTO

Ya hemos hecho notar la diferencia entre estas dos clases principales de periféricos. Los periféricos de E/S existen fundamentalmente para llevar información dentro o fuera del sistema computador. Usualmente la información es llevada hacia o desde el usuario. Hemos excluido los dispositivos de comunicaciones, los cuales pasan mensajes hacia o desde otro computador, y también hemos excluido los sensores y actuadores usados en procesos de control y que se verán brevemente al final.

Por otra parte, los periféricos de almacenamiento son usados para guardar información que será usada de nuevo por el sistema computador. El medio de almacenamiento (usualmente disco o cinta) está dentro del dispositivo periférico, y no es necesaria la intervención del operador; la computadora tiene un control completo sobre él. Sin embargo, cuando el medio es removible, puede ser desmontado por el operador y almacenado en cualquier parte, y por lo tanto el mismo dispositivo puede almacenar información adicional en nuevas unidades.

Hemos de señalar también cuando el volumen es retirado o separado del periférico pudiendo ser cargado en otro periférico de modo que sirve para diferentes sistemas computadores. En ese caso cada uno de los periféricos actuantes, tan lejos como sea de su propio sistema es concebido como un dispositivo de E/S. Esto es particularmente significativo en el caso de programas, los cuales están casi siempre en cinta o disco magnéticos, y más recientemente en disco óptico.

Hay otras maneras de dividir el espectro de dispositivos periféricos, y una de las más importantes (para programadores e ingenieros, no para usuarios finales) es emparejar fielmente la división entre periféricos de E/S y almacenamiento. En este caso se tiene en cuenta el tamaño de la unidad en la cual los datos son transferidos entre la unidad básica y el periférico y las dos clases son la transferencia de bloques y de caracteres.

Cuando usamos una máquina de escribir, cada uno de los caracteres es transferido al papel tan pronto como presionamos la tecla correspondiente. Usualmente, es lo mismo que cuando usamos el teclado y la pantalla del ordenador, aunque de hecho, esto no significa que sea tan sencillo como parece. En cada caso los datos son transferidos carácter a carácter, cada transferencia es una operación separada, pudiéndose teclear muchos o pocos caracteres, como escogamos.

Cuando los datos son intercambiados entre la unidad básica y los discos o cintas, la transferencia no se realiza de la misma forma. En este caso la unidad básica no transfiere un carácter cada vez, sino que cada operación involucra la transferencia de un bloque de datos, formado con un conjunto de caracteres o bytes. En el caso de discos, la unidad puede ser descrita como sector; hay una diferencia de hecho entre bloques y sectores, pero no necesitamos concentrarnos en esto ahora. Habitualmente, aunque no es necesario, todos los bloques de un mismo medio son de la misma longitud, y la longitud típica de los bloques es 512 bytes. De ésta forma, tanto si el usuario (o más bien el sistema operativo) sólo necesita transferir un carácter o varios, la transferencia es de un bloque completo. Si no hay datos suficientes para completar el bloque, el resto de las localizaciones de carácter son llenados con caracteres, los cuales pueden ser reconocidos como tal, pero no tienen otro significado. Si por el contrario, la información a transferir excede el tamaño de un bloque, se dividirá en tantos bloques como sean necesarios para transferirlos luego en secuencia de forma iterativa. En este caso el último bloque puede quedar incompleto y se necesitarán caracteres de relleno.

En general, los periféricos E/S transfieren datos en caracteres mientras que los periféricos de almacenamiento transfieren en bloques aunque esto no es una regla invariable. En particular, discos y cintas usan casi invariablemente transferencias de bloques aun cuando sean usados para transferir datos dentro o fuera del sistema.

Existen varias razones por las que los dispositivos se diseñan para transferir datos en bloques en lugar de caracteres individuales pero como norma general se usa transferencia de bloques cuando hay movimiento continuo del medio hacia o desde el cual se transfiere el dato. Un ejemplo simple es el de la cinta magnética. El movimiento de la cinta no puede comenzar o parar instantáneamente. En un dispositivo típico la cinta se mueve normalmente más de un centímetro hasta alcanzar su velocidad de trabajo y otro tanto mientras se para. Así, si hemos escrito caracteres individualmente podríamos escribir solamente un carácter por pulgada mientras que una cinta magnética es capaz, de hecho, de almacenar varios miles de caracteres por pulgada. Para usar

una cinta eficientemente una vez que hemos comenzado a moverla debemos escribir muchos caracteres antes de que se pare. Esto se hace por medio de un espacio de almacenamiento dedicado llamado 'buffer' (memoria tampón o memoria intermedia) en el que se recogen los caracteres suficientes para formar un bloque. Se escriben todos los caracteres tan rápido como el dispositivo permita y la cinta se para cuando el 'buffer' se vacía. Esto es una versión simplificada de la forma en que se usa en la práctica una cinta. Así, la cantidad de datos que nosotros transferimos en una operación es un bloque, y aunque el 'buffer' tiene un tamaño fijo es obviamente más eficiente (aunque no esencial) si se hace la longitud de cada bloque igual al tamaño del 'buffer'. En el caso de los discos, la situación es similar aunque los platos giratorios estén rotando constantemente. Para acceder un dato concreto, la cabeza debe realizar una secuencia de movimientos, que permitan localizar primero el fichero a través de la jerarquía de ficheros y luego el carácter concreto dentro del fichero. Esto implica una secuencia de pasos que involucran dispositivos mecánicos y por lo tanto son lentos con respecto a las transferencias puramente electrónicas. Resulta por lo tanto más conveniente la lectura o escritura de bloques completos.

Esta memoria intermedia o buffer puede estar constituida por una memoria físicamente independiente aunque lo más habitual es que una parte de la memoria principal de trabajo se reserve para esta tarea. Si se emplea esta segunda posibilidad, la reserva de memoria así como las transferencias entre ésta y la memoria principal o el dispositivo es gestionada por el sistema operativo. Cuando se trata de una memoria físicamente independiente, reside normalmente en el propio periférico y es gestionada por el hardware del mismo como una cola o memoria FIFO (Primero en Entrar, Primero en Salir).

Existe otra importante razón para transferir datos en bloques y es el tratamiento de errores, tanto para detección como para su posible corrección. Al realizar la recuperación de datos de un determinado dispositivo, pueden haberse perdido o dañado debido a algún defecto en el medio de almacenamiento. Existen métodos para codificar redundantemente datos, en otras palabras, con más bits que el mínimo número necesario para transportar la información, que permiten recuperar datos que se han perdido. Estos métodos sólo son efectivos cuando los caracteres se han transferido en bloques.

Periféricos de almacenamiento

2.1 INTRODUCCIÓN

Los periféricos de almacenamiento tienen varios usos. El más importante es como medio de almacenamiento masivo, donde se almacenan datos y programas, y necesitan estar directamente accesibles al ordenador. Esto tiene un doble propósito, por una parte, reducir la capacidad necesaria en la memoria principal del ordenador, que es cara, y por otra garantizar la retención de la información cuando cae la alimentación del sistema, en cuyo caso el contenido de la memoria principal, se pierde. La segunda función de los periféricos de almacenamiento es el de almacenar una copia de salvaguarda o respaldo ("back-up"), es decir, hacer una copia duplicada de los datos del medio de almacenamiento masivo, para seguridad. Los dispositivos de almacenamiento para esta segunda función tienen medios intercambiables, de muy alta capacidad y de coste reducido. El principal inconveniente de este tipo de dispositivos es su velocidad de acceso. Los periféricos de almacenamiento con medios separables se pueden usar también para permitir la entrada de programas al sistema, y la transferencia de datos de un sistema a otro. Esto requiere que ambos sistemas puedan manejar el mismo tipo de medios, y puedan entender la información grabada en él. Existen formatos estándar para asegurar esto (formatos definidos normalmente por el sistema operativo).

Los periféricos de almacenamiento difieren de la memoria principal del ordenador, especialmente en la no volatilidad (no necesitan alimentación para retener la información almacenada), tienen costes más reducidos por megabyte almacenado, y considerablemente más lentos en términos de tiempo de acceso. Se alcanza este bajo costo, usando medios de almacenamiento en forma de superficies bidimensionales continuas. Uno de los problemas de este tipo de medios es que no disponen de celdas predefinidas para almacenar los datos, y esta debe incorporar también la información necesaria para poder distinguir las distintas celdas de almacenamiento durante los procesos de lectura. Se dispone un pequeño número de puntos de acceso, o cabezas (a menudo una sola); la cabeza o la superficie, o ambas, están en movimiento para hacer coincidir la cabeza y el dato requerido en un mismo punto; por este motivo, el término de almacenamiento dinámico se usa algunas veces. Esto significa que los periféricos de almacenamiento, a diferencia del almacenamiento principal, son dispositivos electromecánicos con partes en movimiento, lo que les hace menos fiables que los puramente semiconductores. Existen dispositivos de almacenamiento no volátil de naturaleza puramente electrónica o basada en semiconductor y con unos tiempos de acceso inferiores pero su elevado coste restringe notablemente su rango de aplicaciones. Otro problema, es que no es posible asegurar que el medio

de almacenamiento esté completamente libre de defectos sin que el coste se vea fuertemente incrementado. Por esta razón, todos los periféricos de almacenamiento hacen alguna previsión para detectar errores en los datos almacenados, y también para corregir estos errores.

La mayoría de los dispositivos de disco se basan en la tecnología de grabación magnética. Después de más de 40 años de desarrollo esta tecnología puede considerarse, al menos en sus fundamentos, como una materia estable y madura. En todo almacenamiento magnético se distinguen dos procesos: lectura y escritura.

Todo medio de almacenamiento, se basa en la alteración de alguna propiedad de un medio material. Si esta alteración es reversible el medio puede almacenar distinta información en distintos momentos. Si no lo es, tan sólo se podrá grabar una vez y la información permanecerá en el medio en el futuro. La propiedad que emplean los dispositivos magnéticos, es la orientación de los dominios magnéticos de un material ferromagnético. Los materiales ferromagnéticos tienen la particularidad de que los dominios magnéticos, que son las porciones de material más pequeñas que tienen una misma orientación de su campo magnético, pueden ser orientados por un campo magnético externo y mantienen esta orientación cuando el campo magnético desaparece. En esto se distinguen de los medios diamagnéticos en los que los dominios vuelven a sus posiciones iniciales cuando el campo externo desaparece. Si posteriormente se aplica un campo magnético en otra dirección los dominios o dipolos magnéticos se reorientan de acuerdo al nuevo campo. Es por tanto un proceso reversible y como consecuencia los materiales magnéticos pueden alterarse tantas veces como se desee.

2.2 TAMBORES Y DISCOS MAGNÉTICOS

Los medios de almacenamiento masivo han sido una parte esencial de los sistemas computadores, desde el inicio de estos. El primer medio de almacenamiento masivo que se usó, fue el tambor magnético, el cual desembocó en los discos magnéticos, que han sido el pilar fundamental del almacenamiento masivo de los ordenadores desde los años sesenta. En los noventa, los almacenamientos magnéticos están sufriendo un serio cambio hacia el almacenamiento óptico. En la actualidad, los discos magnéticos son aún los dispositivos de almacenamiento más usados, y de hecho, con el teclado y la pantalla, el más común de todos los periféricos. La importancia de este tipo de dispositivos está avalada por varias razones:

- ♦ Todos los sistemas informáticos disponen de algún disco duro.
- ♦ Es uno de los sistemas de almacenamiento más experimentado, debido en parte a ser de los más antiguos.
- ♦ Presentan un compromiso interesante entre coste y prestaciones.
- ♦ Tienen una influencia considerable en las prestaciones globales del sistema completo.
- ♦ Es uno de los puntos más críticos y más débiles del sistema.

Las características básicas que tiene un medio de almacenamiento masivo son: no volatilidad, y menor coste en comparación con la memoria principal de la computadora. Los tiempos de acceso son mucho mayores que los de la memoria principal, a veces se vuelven demasiado grandes y limitan el rendimiento del sistema. La transferencia de datos hacia y desde la unidad básica debe ser razonablemente rápida con velocidades que varían entre 10 Mbytes y 80 Mbytes por segundo e incluso superiores en sistemas de última generación. Estos parámetros se consideraban bastante buenos, hasta que apareció un serio competidor en los primeros discos ópticos, surgidos en el mercado a finales de los 80.

Los almacenamientos magnéticos son intrínsecamente no volátiles, y las tecnologías de grabación magnética han ido evolucionando desde las cintas utilizadas en las primeras

computadoras, que tenían tiempos de acceso bastante elevados (hasta de minutos). Los cortos tiempos de acceso requeridos por la unidad básica, hacen necesario ir directamente al dato buscado sin que se tenga que pasar por otros (como sucede en la cinta, en la que el acceso es secuencial). Para estos requerimientos, las formas convenientes son los discos y tambores, y los dispositivos que utilizan esta forma de medio se describen como dispositivos de almacenamiento de acceso directo o DASD ('Direct Access Storage Device').

La distinción entre RAM ('Random Access Memory'), y el DASD, es que el primero accede de forma inmediata a cualquier byte de datos requerido, mientras que el DASD accede sólo a un bloque que contiene típicamente un Kbyte de datos, más o menos. El bloque se transfiere entero a la RAM, y sólo entonces la unidad básica puede acceder a bytes específicos. El acceso a los bloques sin embargo es totalmente aleatorio pudiendo acceder a cualquiera de ellos sin tener que pasar por otros. No obstante, este acceso a los bloques de un determinado fichero no es totalmente directo puesto que el sistema debe localizar primero la ubicación del archivo y su distribución sobre el soporte magnético. Para poder acceder a esta información el sistema operativo debe consultar la estructura jerárquica de su sistema de archivos, que obviamente también estará almacenada en el medio magnético, pero una vez consultada esta tabla y localizado el fichero, el acceso a un bloque concreto puede hacerse de forma directa sin necesidad de recorrer otros bloques del mismo. Esta situación no es posible en el caso de las cintas magnéticas de cualquier tipo, donde para que la cabeza de lectura alcance un determinado bloque dentro del fichero ha tenido que recorrer todos los bloques anteriores aunque no sean leídos y transferidos al sistema principal.

Hay también varias formas cualitativas, en las que se pueden clasificar los DASD. Una forma de distinguir los dispositivos es clasificándolos en medios flexibles y rígidos; otra es entre dispositivos en los que se puede intercambiar el medio y en los que éste permanece fijo en su sitio. Normalmente, los discos realizados con material rígido son fijos, por lo que se les ha dado en llamar discos duros, aunque no tardaron en diseñarse los discos extraíbles con material rígido. Junto a los discos rígidos, se desarrolló otro tipo de discos que utilizaban medios flexibles, típicamente llamados disquetes ('floppy disk').

2.3 CABEZAS POR PISTA Y MOVIMIENTO DE LAS CABEZAS

Un rasgo fundamental en el diseño de los dispositivos de acceso directo, en contraste con los de acceso secuencial, es que los datos están almacenados en un gran número de pistas separadas, cada una de las cuales almacena sólo unos pocos Kbytes de datos. Los accesos rápidos se consiguen permitiendo que se acceda a cualquier pista, escrutándose los datos secuencialmente. Los primeros dispositivos de acceso directo en los que ocurría esto antes que los discos eran los tambores, y aparecieron en dos versiones. En la primera versión tenían cabezas separadas para cada una de las pistas, a menudo varios cientos de cabezas. Esto se dio en llamar tambor de cabeza por pista. En otra versión, una cabeza simple, o un pequeño grupo de cabezas, podía moverse paralelamente al eje del tambor, para enfrentarla a una pista o grupo de pistas. Esto se llamó tambor de cabeza móvil.

2.4 TAMBORES Y DISCOS

Los primeros dispositivos de almacenamiento de acceso directo, se fabricaron en forma de tambor, debido en parte a la facilidad de fabricación, y en parte para que todas las pistas fuesen idénticas con lo que se simplificaba el diseño. Cuando se incrementaron las necesidades de almacenamiento, se hizo difícil mantener los cilindros de tamaño razonable, además la deposición de la película magnética en la superficie plana de un disco es mucho más sencilla y fiable que

sobre la superficie curva del cilindro. Cambiándose a la forma de disco (con el gramófono como precedente), fue posible un diseño más compacto, usando ambas superficies del disco. Esto también facilitó la aparición de medios intercambiables, al menos en los dispositivos con cabeza en movimiento. El siguiente paso lógico, fue el usar varios discos. Los dispositivos de disco fijo con una cabeza por pista (HPT, 'Head Per Track') se han estado usando hasta hace poco, aunque para distinguirlos nos referimos a ellos como tambores, para distinguirlos de los dispositivos de cabeza en movimiento, a los cuales se les aplica ahora el termino universal de unidades de disco. Algunos sistemas combinan dos técnicas; la sección principal con cabeza en movimiento, y una sección suplementaria con un pequeño número de cabezas fijas para usarlas donde se necesitan unos accesos mucho más rápidos.

Las pistas de los discos magnéticos están dispuestas en círculos concéntricos. En consecuencia, todas las pistas no son iguales, ya que el diámetro de cada una difiere del de las demás. Aunque podrían emplearse otras soluciones (como en los discos compactos), todos los discos magnéticos modernos giran a una velocidad constante, y escriben y leen datos a una velocidad constante. La longitud de la pista ocupada por cada uno de los bits de datos, varía por tanto, de una pista a otra. En la práctica, la pista utiliza una parte relativamente limitada de la superficie del disco, el radio de la pista más interna está entre la mitad y las dos terceras partes del radio de la pista más externa, pero esto todavía requiere una gran tolerancia en el sistema de lectura/escritura. En muchos discos, uno o más parámetros del disco se cambian con el radio variable de la pista que se comienza a acceder; normalmente sólo la corriente de escritura que determina la fuerza del campo magnético usado para escribir datos en el disco. Otra modificación consiste en la precompensación que se describirá más adelante.

2.5 DISCOS Y CABEZAS MÚLTIPLES

Ya hemos visto que las razones principales por las que cambiamos del tambor al disco, son la disponibilidad de uso de ambas caras del disco y de discos múltiples. En principio, podría ser posible el empleo de una sola cabeza de lectura/escritura, y moverla de superficie a superficie. Esto no sería una solución práctica, puesto que el coste y el bajo rendimiento aparecerían como factores negativos. Todas las unidades de disco magnético tienen cabezas separadas (algunas veces más de una), para cada una de las superficies de grabación (Fig. 2.1 a). En la figura (2.1 b) se muestra la imagen del interior de un disco duro donde se aprecian varios platos circulares junto con los brazos que soportan y desplazan las cabezas. Por otra parte, a excepción de algunos pocos dispositivos especializados, hay un único canal de datos, que se conecta a la cabeza requerida mediante un árbol de multiplexores electrónicos.

De esta forma, sólo se usa una cabeza cada vez, por lo que no es necesario mover cada una de las cabezas por separado, y se mueven todas juntas para situar una sola cabeza. Estas, se sitúan al final de una serie de brazos, de modo que cada brazo alcanza a pares de superficies adyacentes. Es decir, al final de cada brazo hay un par de cabezas, una por cada superficie adyacente, excepto el brazo adyacente a las superficies más externas, que tienen una sola cabeza cada uno. En unidades con discos intercambiables, la superficie más externa no se usa, porque el riesgo de daño es grande, excepto cuando los discos están permanentemente cerrados en la carcasa. Siempre habrá una pista (una por cada superficie de grabación), que está enfrentada a la cabeza correspondiente en cierta posición, y por tanto, accesible simplemente conmutando sin ningún movimiento de cabeza; el conjunto de pistas de todas las superficies que simultáneamente están enfrentadas a las distintas cabezas es lo que se llama cilindro. Es decir, la 'superficie cilíndrica' está formada por un número de pistas idénticas y dispuestas verticalmente sobre cada uno de los discos (Fig. 2.1 a). El número de cilindro es una de las tres componentes de dirección necesarias para encontrar una dirección específica. Las otras dos componentes son el número de cabeza y el número de sector. Hay que tener en cuenta que un disco con más de una superficie de almacenamiento, que es lo

habitual, tiene una estructura tridimensional, por lo que se requieren tres coordenadas para acceder a un determinado elemento de información.

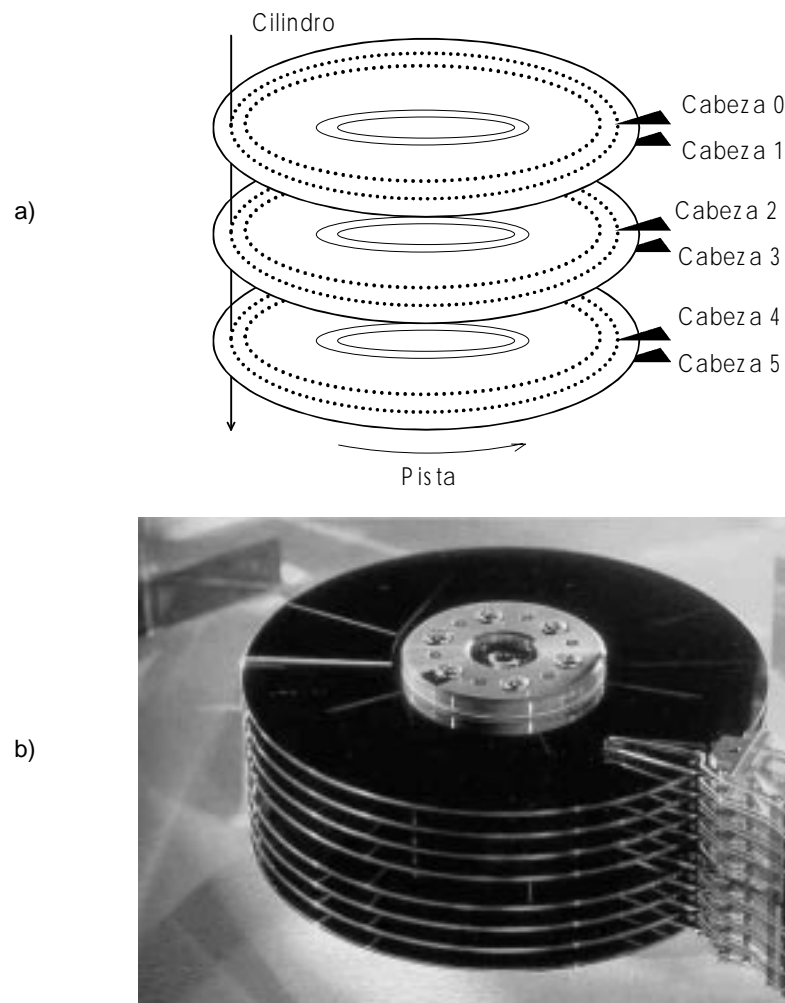


Fig. 2.1 Esquema de un disco con múltiples platos y cabezas (a). Imagen que muestra el interior de un disco duro donde se aprecian los distintos platos y los brazos que soportan y desplazan las cabezas (b).

El tiempo de acceso entre bloques de datos dentro del mismo cilindro, es más corto que el acceso a los datos de otros cilindros, porque las cabezas no necesitan desplazarse, aunque sí hay que tener en cuenta el tiempo de latencia que es el tiempo que transcurre desde que la cabeza se posiciona en la pista o cilindro adecuado hasta que el sector a leer se enfrenta con la cabeza. Este tiempo tiene una cota máxima que es el tiempo que tarda el disco en dar una vuelta completa, ya que durante este periodo todos los sectores pasan por delante de la cabeza de lectura/escritura. El tiempo de latencia es más bajo que el de posicionamiento sobre la pista, ya que al estar girando constantemente no se precisa vencer la inercia. Este tiempo de latencia viene dado únicamente por la velocidad de giro del disco y en principio interesa que sea lo más alta posible. Sin embargo, esto tiene varios problemas: por una parte supone un mayor desgaste de los rodamientos, un mayor consumo y un calentamiento mayor y por otra exige una velocidad de transferencia más alta lo que implica un diseño del canal de transferencia mucho más costoso. Como tiempo medio de latencia se considera el tiempo que tarda en completar $1/2$ vuelta. De esta forma un disco que gira a 7200 rpm. tiene un tiempo de latencia medio de 4.2 ms.

Cuando el tiempo de acceso es importante, cada superficie puede incorporar dos o incluso más cabezas. En este caso existe igualmente un movimiento de cabezas, pero hay espacios separados radialmente, de forma que cada una de las cabezas usa la mitad de las pistas. El árbol de

multiplexores se extiende para permitir las cabezas extras; el efecto es el doble número de pistas por cilindro, y dividimos el número de cilindros. El posicionamiento de la cabeza, ahora, sólo necesita moverse en la mitad del rango original, y la media de movimiento queda dividida por dos. De esta forma, el tiempo medio de búsqueda (tiempo en mover la cabeza a la pista requerida), y el número de búsquedas (p. ej. cambios de cilindro), se reducen ambos, aunque no existe reducción en la latencia. El problema de la variación del espaciado de bits entre pistas, es de más fácil solución, ya que cada una de las cabezas se puede optimizar para su propia selección de pistas. En algunos discos de cabezas móviles de alto rendimiento hay dos posicionadores de cabezas separadas, donde cada uno tiene acceso a la mitad de las pistas en cada una de las superficies, uno a la mitad interna y el otro a la mitad externa. Esto, reduce de nuevo el tiempo medio de acceso, así como el número de movimientos de cabeza necesarios, aunque la extensión de esto varía mucho con la aplicación. Muchos actuadores incrementan bastante el coste de la unidad, el uso de dos unidades cada una con la mitad de capacidad suele ser la mejor solución. La única forma de reducir la latencia, es proporcionar dos cabezas por pista, diametralmente opuestas, una a otra. Esto obliga a dos actuadores independientes, ya que si se situasen en el mismo brazo, cuando una de las cabezas se estuviese moviendo hacia el centro, la otra lo haría hacia el exterior y no estarían simultáneamente sobre la misma pista. Otra forma de reducir el tiempo de acceso, es proporcionar una memoria caché, que podemos considerar como un 'buffer' muy grande. No obstante, ésto queda normalmente fuera del ámbito de los discos magnéticos siendo una tarea más propia de los distintos sistemas operativos.

2.6 ESPACIADO ENTRE CABEZAL Y DISCO

Los discos flexibles, normalmente giran en contacto con la cabeza de lectura/escritura, al igual que las cintas magnéticas. La densidad de datos de estos discos es baja, lo que permite tener una espesa capa magnética en el disco, y una cabeza robusta. El disco gira relativamente lento (300 r.p.m.), y cuando los datos no se leen ni escriben, la cabeza se retira del contacto con el disco y éste deja de girar. Por tanto, aunque haya algún desgaste de la cabeza o del disco, si éste es pequeño, no importa demasiado ya que sólo se produce cuando hay lectura o escritura.

Este no es el caso de los discos duros y tambores, en los que hay una capa más delgada y cabezas más pequeñas, y giran a unas 3600 r.p.m. y algunos más modernos a 7200 e incluso más. Estos discos están girando constantemente ya que al ser su masa mayor tienen una gran inercia lo que hace que se incremente el tiempo que tarda en alcanzar la velocidad estacionaria de trabajo. Aquí es necesario evitar el contacto entre la cabeza y la superficie de grabación ya que el rozamiento es constante. Al ser la densidad de grabación mucho mayor, la capa magnética debe ser mucho más fina lo que origina unos campos magnéticos más débiles y por lo tanto la distancia entre la cabeza y la superficie debe ser muy pequeña y constante. Esto es debido a que las líneas de fuerza del campo magnético tienden a abrirse con la distancia, en otras palabras, la intensidad del campo disminuye y además se amplía la zona de influencia del campo con lo que la zona magnetizada se hace mayor lo que impediría el aumento de la densidad. Para resolver este problema, se desarrollaron varias técnicas con objeto de reducir y mantener constante la distancia entre la superficie y la cabeza (4 micras es un valor usual). La primera consistió en que la cabeza se acercaba a la superficie mediante un tornillo hasta que rozase con la superficie del disco, y a continuación, se aflojaba un poco este tornillo. Esta técnica tuvo poco éxito, y pronto surgió otra técnica derivada del comportamiento del aire, en la que la separación de la cabeza depende de la forma de la cabeza, y de la capa de aire existente entre la cabeza y la superficie del disco (Fig. 2.2). De esta forma, esta película de aire empuja a la cabeza hacia arriba, mientras que un resorte que soporta a la cabeza, empuja hacia abajo, llegándose a un equilibrio entre ambas fuerzas bastante cerca de la superficie manteniendo constante la distancia entre cabeza y disco. Dicha distancia ha ido decreciendo gradualmente, hasta que en los discos actuales se ha llegado a la aproximación anteriormente mencionada.

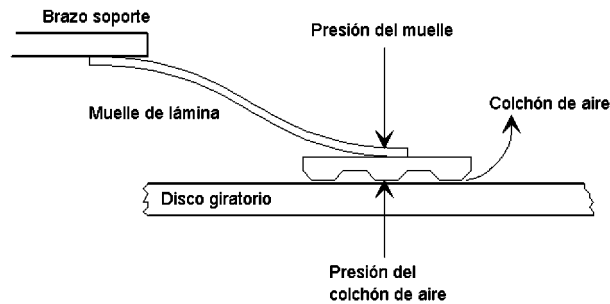


Fig. 2.2 Cabezal 'volando' sobre el disco

En este vuelo de cabezas surgen dos problemas: el primero de ellos está asociado al hecho de que el vuelo sólo se mantiene cuando el disco gira a su velocidad normal. Si el disco disminuye su velocidad, la altura del vuelo va decreciendo y finalmente la cabeza tocará la superficie del disco empujado por el soporte del disco. Existen algunas técnicas para evitar esto, como por ejemplo quitar el soporte que empuja a la cabeza cuando el disco pierde velocidad. Este mecanismo es complejo, y por lo tanto caro. Otro método utilizado consiste en llevar la cabeza a un lugar del disco donde no haya datos (aparcas la cabeza). Indudablemente hay un desgaste de la cabeza, el cual es mínimo, ya que sólo se produce durante el arranque y la parada del disco, es decir los momentos en los que el disco no gira a su régimen normal. Este aparcado de cabezas debe realizarse en una zona no destinada a datos, ya que aunque la cabeza pueda soportar el ligero desgaste del rozamiento de arranque y parada, la delicada película magnética sí podría dañarse, ya que al contrario que en los discos flexibles, no está recubierta por una capa protectora. Esto obliga a retirar la cabeza hacia el interior o el exterior del disco cuando se corta la alimentación. Esto se consigue con un resorte que la alimentación mantiene desactivado y al fallar ésta, automáticamente empuja la cabeza hacia uno u otro extremo de su recorrido antes de que el disco deje de girar completamente. Hay que tener en cuenta que una vez que se corta la alimentación, el disco sigue girando durante algún tiempo debido a su inercia y al bajo rozamiento que presenta.

El otro problema existente con el vuelo de los cabezales ocurre cuando en la superficie del disco existen rugosidades, o contaminación debido a las impurezas del aire, tales como polvo, ceniza de tabaco, etc., lo cual obliga a tener una serie de prevenciones en el almacenamiento del disco.

La solución introducida por IBM fue el 'winchester', en el que el disco y la cabeza se ensamblan juntos en un recinto cerrado que no vuelve a abrirse nunca más. Esto significa que los discos 'winchester' intercambiables no sólo son los discos, sino también incluyen las cabezas y mecanismos de movimiento de éstas.

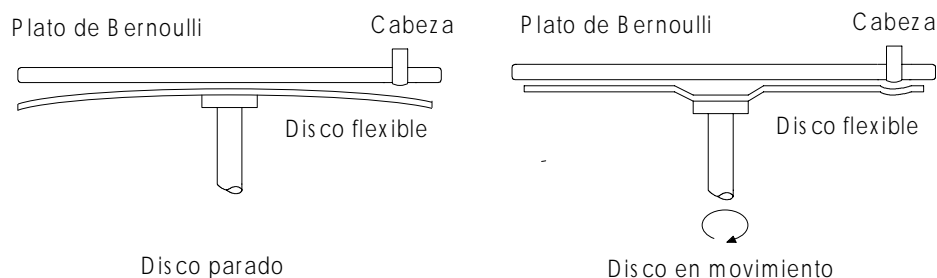


Fig. 2.3 Disco de tipo Bernoulli en funcionamiento.

Aunque ya en desuso, han existido otro tipo de dispositivos que combinan alguna de las características de los discos rígidos y flexibles en el dispositivo, usando el principio de Bernoulli (Fig. 2.3). En estos dispositivos, un disco flexible gira paralelo al piso de metal. Una delgada

película de aire se forma entre los dos, y el espesor de ésta es estable. Como acabamos de describir, las cabezas de lectura/escritura vuelan sobre el disco rígido. El disco también volará sobre cualquier pequeña protuberancia del 'plato', formando pequeños rizos. Esto, no sólo significa que pasará, sin producir daño, sobre el polvo, sino que también sobre las cabezas de lectura/escritura que se mueven radialmente sobre el plato.

2.7 DISEÑO DE CABEZAS

La cabeza es un pequeño dispositivo que lee y escribe los datos en el medio magnético. Durante la escritura de datos, pulsos eléctricos enviados a la cabeza crean áreas magnéticas en el medio orientando los dominios magnéticos del material en uno u otro sentido en función del sentido de la corriente enviada a la cabeza. Durante la lectura, estas áreas magnéticas crean pulsos eléctricos en la cabeza. Para ser más precisos, los pulsos son creados por la transición o el paso de la cabeza de una zona magnetizada en un sentido a otra magnetizada en sentido contrario, ya que un campo magnético constante no es capaz de inducir ninguna corriente. Para que se induzca una corriente es preciso un cambio en el campo magnético. Ver figura (2.9).

Los vuelos de las cabezas es lo más difícil de diseñar en una unidad de disco. Hay requerimientos eléctricos, magnéticos, mecánicos y aerodinámicos y algunos de ellos entran en conflicto entre sí.

Fundamentalmente, las cabezas convencionales de grabación consisten (Fig. 2.4) en un anillo o núcleo de material con una baja reluctancia magnética, con un estrecho hueco cortado en él que constituye el entrehierro (la reluctancia puede considerarse, en el campo magnético, el equivalente a la resistencia).

Se coloca un arrollamiento conductor en el núcleo, de tal forma que cuando pasa la corriente a través de la bobina se produce un campo magnético. Si no hubiese aire en el núcleo de metal, el campo magnético estaría concentrado dentro del material, puesto que su reluctancia es mucho menor que la de los alrededores. Sin embargo, el aire (o el material magnético) del entrehierro tiene mucha mayor reluctancia que el núcleo del material, y por lo tanto, el campo magnético tiende a esparcirse mucho más (Fig. 2.5). Si un disco o cinta con una capa magnética se coloca muy cerca del entrehierro, algunas de las líneas de fuerza del campo magnético pasarán muy cerca de la capa y pueden cambiar el estado magnético de ésta.

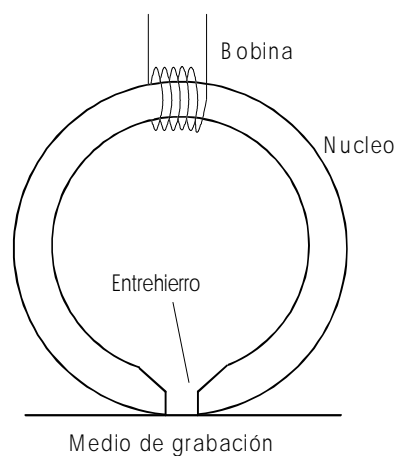


Fig. 2.4 Modelo simplificado de cabeza magnética

A diferencia del núcleo, la capa permanece magnetizada después de que el campo magnético desaparezca debido a la histéresis magnética. La dirección de magnetización depende de la dirección de la corriente a través de la bobina, y esta corriente puede ser reversible. En el disco o cinta, la capa magnética va moviéndose constantemente bajo el hueco del núcleo. Por lo tanto, se produce una secuencia de cambios de flujo magnético sobre ella que se corresponden con los cambios en las corrientes del arrollamiento. Este flujo magnetiza la capa del medio de forma permanente, hasta que se aplique un campo magnético en sentido contrario.

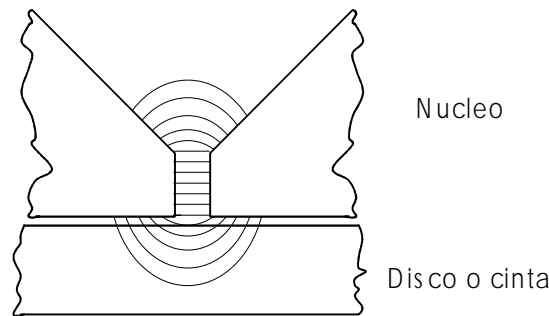


Fig. 2.5 Campo magnético en el entrehierro

Cuando leemos, el campo magnético realmente induce un voltaje en el núcleo, correspondiente a la dirección de la corriente en la que se produjo este campo magnético remanente. El ancho del entrehierro (separación entre los polos) del núcleo (que puede ser menor de un micrón) determina la longitud más corta de la capa que puede ser magnetizada en una dirección, y por tanto, la densidad con la que los datos se pueden empaquetar a lo largo de la pista. El ancho del núcleo en sí mismo, medido perpendicularmente a la dirección del movimiento, es uno de los factores que determina el espaciado de las pistas. Este es considerablemente más grande que la abertura del entrehierro, por lo que la celda de grabación es mucho más ancha que larga. El material del núcleo es usualmente ferrita aunque actualmente se emplean distintos tipos de materiales cerámicos amorfos.

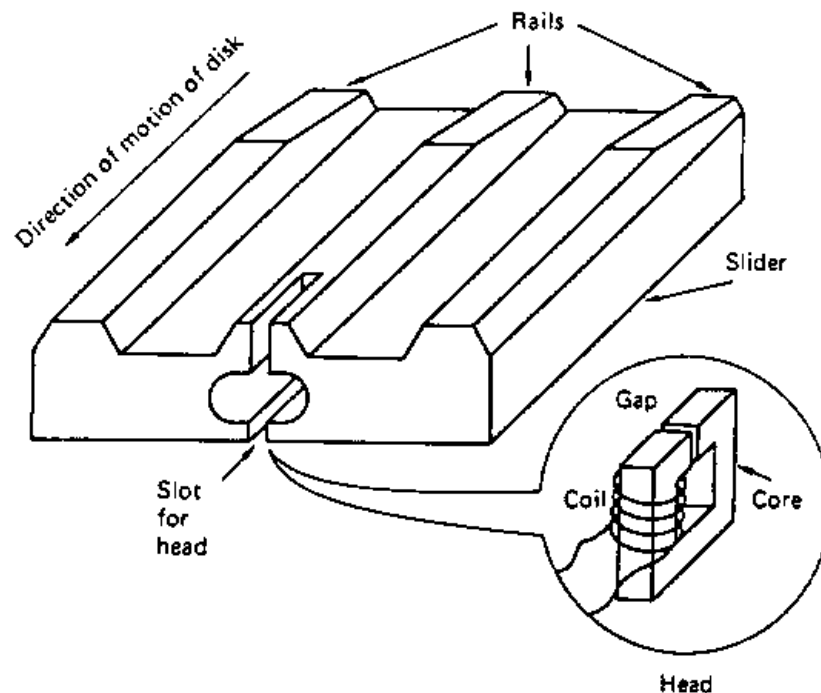


Fig. 2.6 Composición de la cabeza en el cabezal

En la actualidad, existen dos tipos de cabezas: monolíticas y de película delgada. El primer tipo ha sido usado durante muchos años. Consiste en un núcleo de ferrita, en el que se arrolla la bobina (de cable extremadamente fino). Todo el núcleo está diseñado para darle unas propiedades aerodinámicas para que pueda volar a una distancia correcta de la superficie del disco. La mecanización del soporte de la cabeza involucra normalmente tres railes en la cara de la cabeza, paralelos a la dirección en la cual se mueve el disco (Fig. 2.6). La película de aire entre estos railes y la superficie del disco, hace que la cabeza vuele. Dos de estos railes están a los lados de la cabeza. El tercero, en el centro, es el activo y en él está situado el espacio para insertar el núcleo de ferrita que constituye el elemento magnéticamente activo. Una variante de esto, son las cabezas compuestas. La carcasa de la cabeza está hecha de material magnético inerte, y la cabeza de ferrita es mucho más pequeña y está insertada en la carcasa.

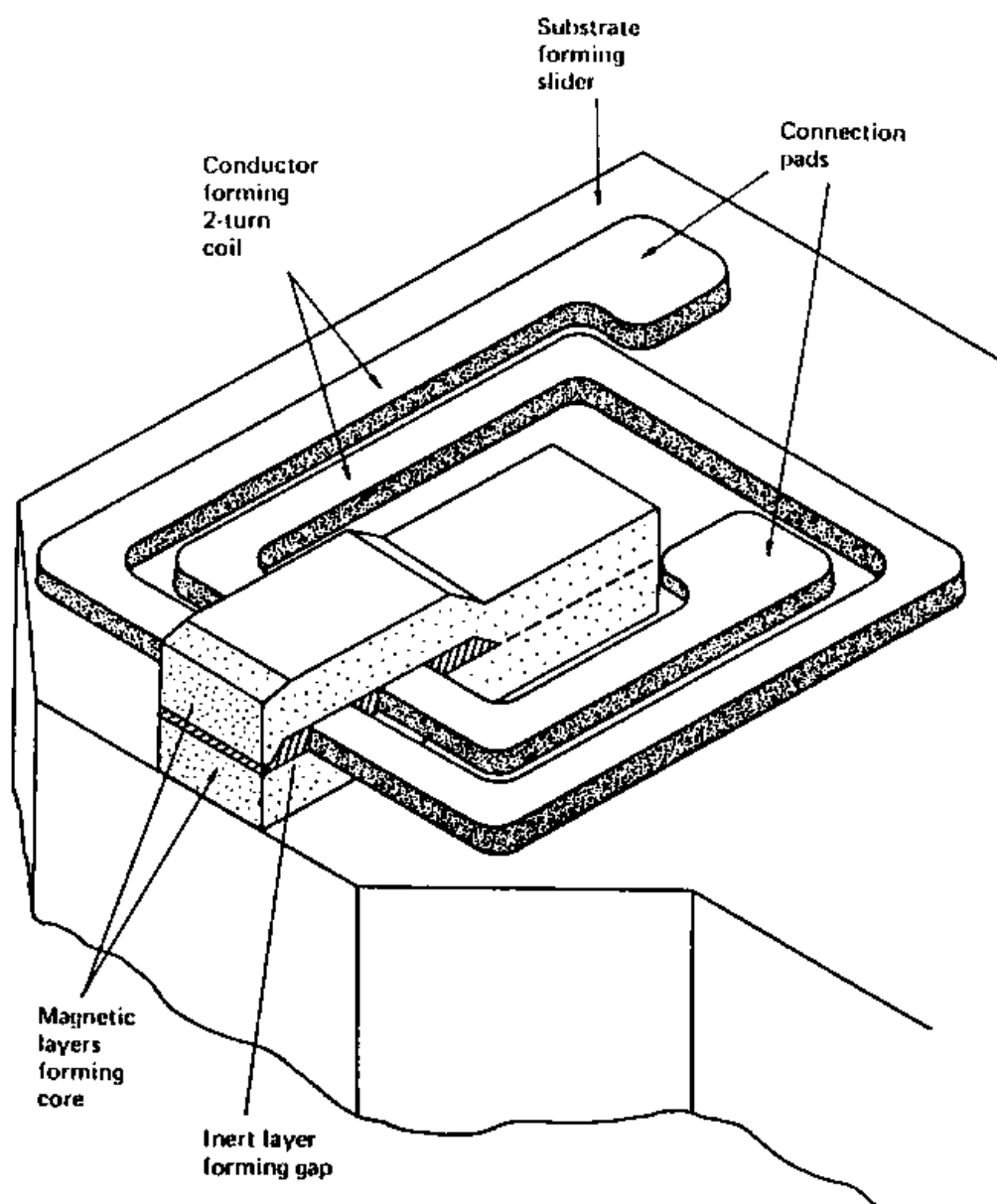


Fig. 2.7 Detalle de una cabeza de película delgada

En las cabezas de película delgada (Fig. 2.7) se usa también un deslizador inerte, pero en este caso, la parte activa de la cabeza está realizada poniendo varias capas en el sustrato, utilizando técnicas similares a las de la industria del semiconductor. El material del sustrato, generalmente

forma una 'carcasa' inerte en el deslizador. Al igual que se hace con los semiconductores, muchas cabezas se fabrican juntas en un único sustrato. La primera cara es una mezcla de metal, seguida de una cara inerte, la cual forma el entrehierro de la cabeza. Una o más caras siguen con el patrón de conducción entre las caras de aislamiento, que constituirán la parte de bobinas con unas cuantas vueltas (normalmente entre 2 y 20). Encima se pone otra mezcla de metal. Esto hace contacto con la primera, pero está separada de él por un hueco relleno. El sustrato es entonces cortado en cabezas individuales, y cada una de ellas es mecanizada con un deslizador para el correcto perfil de vuelo. Estas cabezas, suelen tener dos railes (Fig. 2.8) en vez de tres, y están fabricados con cabezas de película delgada separadas en cada rail. De estas dos cabezas solo se usa la que tiene mejores propiedades después del testeo. Las dos se hacen simplemente para incrementar la probabilidad de que una de las dos sea buena.

Las cabezas de película delgada pueden hacerse con dimensiones más finas y precisas que las cabezas monolíticas. En principio serían menos caras de hacer, aunque el paso de mecanización final es crítico y dificultoso.

2.8 POSICIONAMIENTO DE LA CABEZA

Hemos mencionado hasta ahora que existe un solo mecanismo de posicionamiento, que es el que mueve todas las cabezas a la vez. El diseño de este mecanismo tiene un considerable efecto en el coste del dispositivo completo y en su rendimiento, particularmente en el tiempo de acceso. El mecanismo tiene dos partes: el conjunto de los brazos que lleva las cabezas, y el actuador que controla su posición. El brazo está diseñado para moverse en línea recta, por lo que la cabeza se mueve a lo largo del radio del disco y su eje está siempre tangencial a la pista.

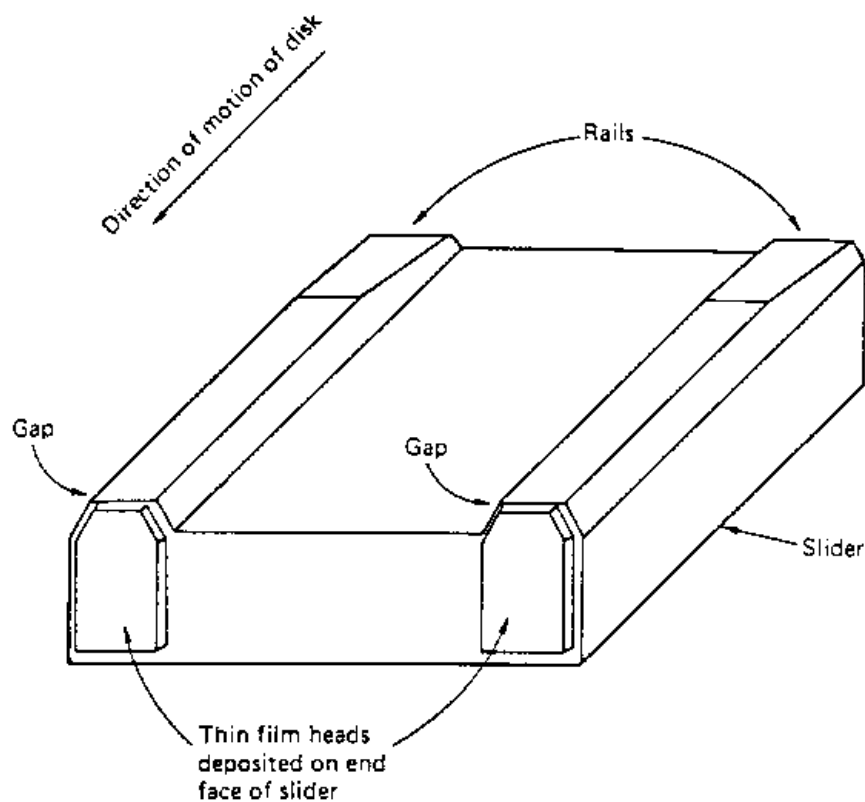


Fig. 2.8 Deslizador con cabezas de película delgada

La parte más crítica del mecanismo de posicionamiento es el actuador. Hay dos tipos básicos. Uno está basado en el motor paso a paso, "stepper motor", el cual es un simple motor que se puede girar a un ángulo definido mediante unos pulsos de corriente en su bobina. El segundo tipo es el actuador electrodinámico o de bobina móvil que se basa en el mismo principio de funcionamiento que los altavoces: esto es un simple bobinado en un campo magnético permanente (imán), como el de un altavoz, de ahí su nombre ("voice coil"). Este último es más rápido, consume mucha menos potencia y permite un mejor posicionamiento al no estar restringido a unas posiciones fijas como el motor paso a paso, pero es más caro y más difícil de controlar.

Asociado con el posicionamiento, debe haber algún método para determinar la posición actual de la cabeza. En las unidades de disco flexible, y en algunos de los discos duros más baratos, esto viene dado por el cómputo o cálculo estimado sin realimentación. La posición de la pista se localiza simplemente contando el número de pulsos aplicados al motor paso a paso, el cual se mueve un número fijo de pasos (a menudo uno) por pista. Esto es sólo satisfactorio cuando la densidad de pistas es relativamente baja, típicamente sobre 400 ó 500 pistas por pulgada. Igualmente, este espaciado sólo se puede conseguir en discos fijos, donde el mismo mecanismo es el que lee y escribe los datos. En los discos flexibles esto no ocurre, puesto que la lectura y escritura de datos se suele realizar en controladores de dispositivo distintos, por lo que la densidad de las pistas es menor. Para una alta densidad de pistas, con un intervalo de 2000 o más pistas por pulgada, es necesario algún tipo de servosistema. Todos los actuadores precisan de algún tipo de servosistema para controlar la posición de la cabeza, salvo algunos tipos de motores paso a paso.

El tiempo empleado por las cabezas para alcanzar la pista requerida y situarse sobre ella se llama tiempo de búsqueda. Es únicamente de unos 50 milisegundos, cuando el actuador es un motor paso a paso y con los actuadores electrodinámicos o de bobina móvil está por debajo de los 10 milisegundos. Este tiempo se mide normalmente, sumando el tiempo necesario para acceder a un elevado número de pistas en orden aleatorio y dividiendo al final la suma total por el número de búsquedas. El tiempo de búsqueda no incluye la latencia (que es el tiempo requerido desde que se sitúa la cabeza del disco en la pista, hasta que se encuentra el sector correcto). Este tiempo se establece como la mitad del periodo de revolución del disco, por lo que para un disco que gira a 3600 r.p.m. este tiempo será de 8.3 milisegundos si el disco tiene un único grupo de cabezas o de la mitad si el disco, tal y como se comentó anteriormente tiene dos grupos de cabezas diametralmente opuestas. Esto es porque el sector puede encontrarse con una cabeza dos veces en cada revolución. El tiempo total de acceso es la suma del tiempo de búsqueda y el tiempo de latencia.

2.9 EL MEDIO MAGNÉTICO

Las capas magnéticas de discos y tambores consisten en una fina capa activa (ej. magnetizable) en un sustrato inerte más robusto. En el caso de los discos flexibles, el sustrato se fabrica en plástico, generalmente poliéster, tal como MYHER. Este es el mismo material utilizado para las cintas magnéticas, pero el sustrato de los discos es mucho más grueso que el de la cinta para que conserve la forma cuando gire. La estabilidad geométrica o dimensional tiende a ser un problema con este material. El disco se expande y contrae un poco, a menudo más en una dirección que en otra, con las variaciones de temperatura y humedad. Esta es una de las razones por las que los discos flexibles tienen unas pistas más anchas, y por lo tanto, de menor capacidad. Al igual que en las cintas, los disquetes están recubiertos de una capa protectora ya que la cabeza está en contacto permanente con el mismo y de lo contrario la película magnética resultaría dañada. Los discos duros, por el contrario, carecen de esta última capa protectora. Los discos flexibles están siempre encerrados en una carcasa protectora. Existen dos tipos de estas envolturas. Las más viejas consisten en una simple envoltura de plástico con una capa de lana en la cara más interna para limpiar y reducir la fricción de la superficie del disco en su rotación. El agujero de esta carcasa

para el acceso de la cabeza no está protegido, por lo que se tiene una superficie vulnerable a huellas. El nuevo tipo es una carcasa de plástico rígido, con una contraventana corrediza, la cual cubre el agujero por el que la cabeza accede a la superficie, cuando el disco está fuera de la unidad de disco. Este tipo de diseño ofrece mayor protección, aunque aumenta un poco el coste. Todos los tipos de discos flexibles y de hecho, todos los discos intercambiables tienen alguna clase de protección contra la escritura, que se utiliza para prevenir la destrucción accidental de datos en el disco y que no puede ser modificable por software.

Los discos rígidos y tambores tienen un sustrato de metal, normalmente aluminio. Algunos modelos experimentales han sido fabricados con sustratos plásticos, más baratos, pero con un coeficiente de expansión varias veces mayor que el del aluminio, por lo que es más difícil conseguir un espaciado de pistas constante. También se ha experimentado con sustratos de cerámica y cristal ya que son menos sensibles a los cambios térmicos. El sustrato debe ser extremadamente liso, debido a que la más mínima rugosidad en la superficie influye bastante en la altura del vuelo de la cabeza.

Otra componente vital del disco es la capa magnética. Debemos tener también información sobre el espesor de esta capa. El espesor de la capa es uno de los factores que más directamente influyen en la densidad de datos del disco: a mayor densidad, menor debe ser el espesor de la capa de material magnético. En general, se usan dos tipos de capas: de óxido y de película delgada.

La capa de óxido consiste, generalmente, en partículas de óxido de hierro. La capa es aplicada en forma líquida. La medida de la cantidad de líquido se pone en el disco cuando éste gira, con lo que el líquido tiende a extenderse por toda la superficie. Esta cara tiende a ser más gruesa en el perímetro del disco, y algunas controladoras de disco compensan esta variación, cambiando la corriente de escritura con el radio de la pista.

La capa de óxido es usada en los discos más económicos. Esta tiene buenas propiedades magnéticas y los discos son relativamente baratos de hacer, pero puede ser difícil conseguir un grosor homogéneo y evitar los llamados "agujeros de alfiler", que son puntos pequeñísimos que no han quedado cubiertos por la película magnética, como si se hubiese tocado esta capa con la punta de un alfiler. La capa de óxido tiene un espesor típico de 20 a 30 micras.

La capa de película delgada es generalmente una aleación de metal. La composición exacta varía de un fabricante a otro, pero las componentes principales son cobalto, níquel y fósforo. La capa es más delgada que la de óxido y puede llegar hasta 2 o 3 micras. Hay dos métodos de fabricación comunes: uno es el 'plating', donde el sustrato se sumerge en un baño electrolítico, en el cual se deposita la capa. El segundo método es el 'sputtering'. En este método, el sustrato es introducido en una cavidad en la que se hace el vacío y en la que hay un cátodo fabricado con la aleación a ser depositada. Una corriente eléctrica provoca que las partículas sean emitidas por el cátodo cuando éste se calienta y se depositen en el disco. Este método es más caro que el 'plating', pero se controla mucho mejor la composición de la película, ya que en el baño electrolítico, conforme se deposita la película baja la concentración de la disolución y esto es difícil de controlar.

La gran mayoría de los discos magnéticos usan una grabación longitudinal. Cada una de las regiones de la capa magnética, la cual representa un bit, es magnetizada en el plano del disco y en la dirección de recorrido de la pista. En todo caso, en el sentido de las agujas del reloj o en sentido contrario según si representa un cero o un uno. Sólo en los límites entre estas regiones, el campo es perpendicular a la superficie. Sin embargo, es posible obtener una mayor densidad de grabación si se usa una grabación vertical, en la que la lámina es magnetizada perpendicularmente a su plano; hacia arriba representa un uno y hacia abajo representa un cero o a la inversa. Sin embargo esta técnica no se usa porque requiere que el disco gire entre los polos de la cabeza que deberían estar perfectamente alineados en ambas caras lo cual no es sencillo con altas densidades de grabación.

Esta técnica sin embargo se emplea en los discos magnetoópticos donde este problema se elimina fabricando unos electrodos más grandes, y empleando un láser para que únicamente el punto iluminado por éste sea alterado magnéticamente como se explicará al final de este tema.

2.10 GRABACIÓN DE PULSOS. PRECOMPENSACIÓN

Cuando un medio magnético ha sido grabado, se ha alterado la orientación de los dominios repartidos por todo él. A partir de este momento, y aunque los medios magnéticos son normalmente bidimensionales, consideraremos que la orientación de los dominios se realiza de forma unidimensional. Es decir, los dominios que pueden considerarse como pequeños imanes, orientan sus polos norte-sur en una única dirección. La información se almacenará en este caso según el sentido de estos pequeños imanes: N-S o S-N a lo largo de la línea de almacenamiento. Para aprovechar la característica bidimensional, se emplean múltiples líneas similares, normalmente de forma circular y concéntricas. Al reorientar los dominios, se dice que hemos magnetizado el material en una determinada dirección. Como la orientación de estos dominios cambia a lo largo de la línea, la curva que representa esta orientación se denomina curva de magnetización. En la figura (2.9) se muestra un ejemplo de una pequeña porción de material magnético con zonas en distintas orientaciones. En esta figura se muestra la curva de magnetización ideal y la curva real. Si la magnetización fuese como la primera curva, las zonas que tienen una determinada orientación N-S o S-N se podrían hacer tan pequeñas como fuese necesario y como consecuencia, se podría almacenar una enorme cantidad de información. Sin embargo, la situación real limita la capacidad debido a esas zonas de transición que obligan a que para cambiar la magnetización del material se requiere un pequeño espacio.

Este espacio depende de numerosos factores siendo los más importantes el grosor de la película, el tamaño de los dominios magnéticos del material, el tamaño de la cabeza, la velocidad de giro y la velocidad de variación de la corriente de escritura.

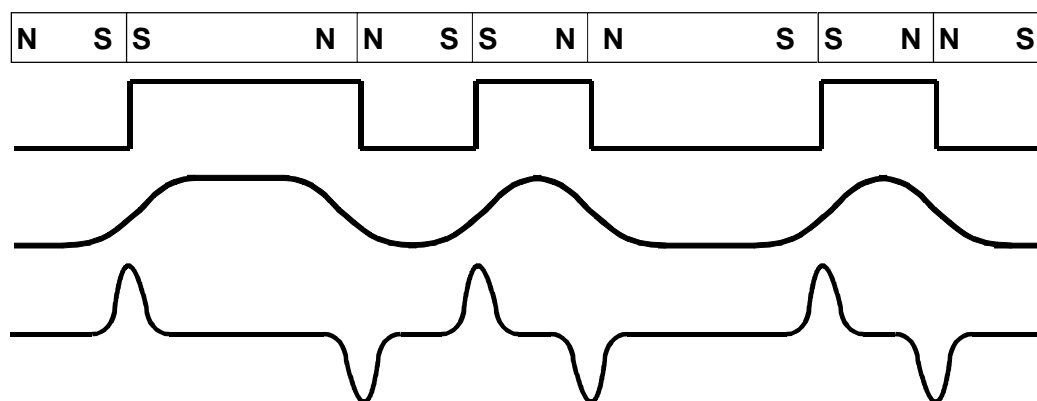


Fig. 2.9 Orientaciones de los dominios magnéticos a lo largo de una línea con la representación ideal de la curva de magnetización y su aproximación real en la que las transiciones no son abruptas. La última curva representa los pulsos de corriente inducida en la bobina de la cabeza durante el proceso de lectura.

Durante el proceso de lectura, hay que tener en cuenta que un campo magnético uniforme no induce ninguna corriente, por lo que la cabeza de lectura únicamente puede detectar los cambios de magnetización y en ese caso se produce un pulso de corriente que puede ser detectado. Este pulso de corriente será de un signo si se pasa de una zona N-S a una S-N y de signo contrario si el paso es a la inversa (de S-N a N-S) (Ver figura 2.9).

Estos pulsos se modelan habitualmente de tres formas distintas:

- Mediante una gaussiana: $e^{-(x-x_0)^2}$
- Mediante la derivada del arcotangente: $1 / (1 + (x - x_0)^2)$
- O mediante el coseno alzado: $\frac{1}{2}(\cos(x - x_0) + 1)$

Estos tres posibles modelos, como es de suponer, son muy similares, de hecho si realizamos su desarrollo en serie de Taylor obtenemos expresiones muy parecidas:

$$e^{-x^2} \approx 1 - x^2 + \frac{1}{2!}x^4 - \frac{1}{3!}x^6 + \dots$$

$$\frac{1}{2}(1 + \cos(x)) \approx 1 - \frac{1}{2 \cdot 2!}x^2 + \frac{1}{2 \cdot 4!}x^4 - \frac{1}{2 \cdot 6!}x^6 + \dots$$

$$\frac{1}{1 + x^2} \approx 1 - x^2 + x^4 - x^6 + \dots$$

Todos los desarrollos incluyen únicamente los términos pares y los signos son alternos. Por este motivo podemos considerar una expresión más general:

$$P \approx 1 - a_2x^2 + a_4x^4 - a_6x^6 + \dots$$

donde los coeficientes a_i de la serie se determinan experimentalmente en el laboratorio.

2.10.1 Superposición lineal. Precompensación.

Como ya se ha comentado, durante el proceso de lectura, la cabeza únicamente detecta las variaciones de magnetización y las convierte a pulsos de corriente que se modelan como se acaba de comentar. Cada pulso en un sentido siempre irá seguido de un pulso en sentido contrario puesto que después de una transición N-S necesariamente debe venir una S-N y viceversa. De esta forma, si los dos pulsos están muy cercanos, se cancelarán parcialmente y habrá que considerar este efecto.

En la figura (2.10) se muestran dos pulsos de signo contrario, con un solapamiento muy ligero y otro par de pulsos con un solapamiento mayor. En este segundo caso se ve que se produce una reducción de la amplitud y un desplazamiento de la posición del máximo y del mínimo respecto de las posiciones que tendrían los pulsos aislados. De estos dos fenómenos, el más grave es el del desplazamiento de la posición de los máximos y mínimos, puesto que es la información que se emplea para sincronizar la lectura y definir de esta forma las distintas celdas. Este efecto se puede corregir parcialmente con una técnica denominada precompensación.

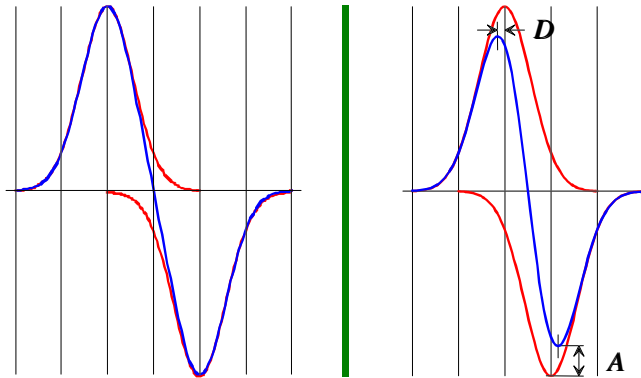


Fig. 2.10 Superposición de pulsos

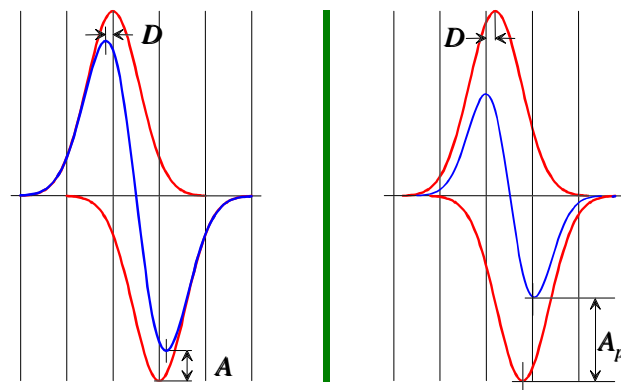


Fig. 2.11 Precompensación de escritura

Con este procedimiento, es posible ajustar las inversiones de flujo cuando están siendo escritas, de forma que la resultante esté en las posiciones en las que estarían los pulsos si no hubiese solapamiento. Las inversiones de flujo que proporcionen un pulso de lectura con un máximo anticipado son escritas más tarde mientras que las que proporcionen un pulso de lectura con un máximo retrasado son escritas de forma anticipada. Este método se ilustra en la figura (2.11). Resulta curioso que para corregir un efecto producido por la superposición estemos planteando más superposición, pero si se observa la figura (2.11) se ve que el desplazamiento de los extremos se ha corregido, pero como contrapartida tenemos una mayor atenuación. Esta mayor atenuación deberá ser corregida con una mayor amplificación. Esto puede amplificar también el ruido por lo que se requerirá también un filtrado. No obstante, esto puede realizarse sólo dentro de unos límites, ya que si la superposición es muy acusada, la atenuación será elevada y los pulsos no podrán ser detectados. De esta forma, el tamaño de una celda magnética, vendrá determinado por el espacio mínimo requerido para albergar una transición que pueda ser detectada.

2.11 OPTIMIZACIÓN DEL ESPACIO. "BANDING"

En relación con la superficie grabable de los dispositivos de almacenamiento magnético deben tenerse en cuenta dos consideraciones previas. En primer lugar la zona interior no se utiliza para almacenar información por tres razones fundamentales: es la zona de soporte del eje, tiene escasa capacidad y la velocidad lineal es excesivamente baja para mantener el vuelo de las cabezas. Por otro lado la zona exterior tampoco se utiliza para almacenar información debido a que es una zona en la que pueden influir notablemente los defectos de mecanizado, así como efectos aerodinámicos de borde y presenta problemas de deposición de la capa magnética si es del tipo de óxido. Por tanto, para calcular el espacio disponible para el almacenamiento debe considerarse únicamente la zona comprendida entre los radios interno y externo, R_i y R_e .

Teniendo en cuenta estas consideraciones, las longitudes del perímetro y las velocidades para una pista interior y una exterior son las siguientes:

$$\begin{aligned} \text{Pista interior:} \quad L_i &= 2\pi R_i, & \text{velocidad} &= V_i = 2\pi\omega R_i \\ \text{Pista exterior:} \quad L_e &= 2\pi R_e, & \text{velocidad} &= V_e = 2\pi\omega R_e \end{aligned}$$

y dado que $R_i < R_e$ entonces $L_i < L_e$ y $V_i < V_e$. Además como es un dispositivo de velocidad angular constante, $\omega = cte$, entonces $Nb_i = Nb_e$ siendo Nb_i y Nb_e el número de bits en las pistas interna y externa respectivamente. Teniendo en cuenta que la densidad de bits por unidad de longitud se define como:

$$B \equiv \frac{Nb}{L}$$

entonces $B_i > B_e$.

Por tanto si queremos mantener una velocidad de datos constante, todas las pistas deben tener el mismo número de bits, a pesar de que las pistas externas serían capaces de almacenar muchos más bits. La máxima densidad de bits está determinada por la pista más interna, que es la de menor perímetro y por lo tanto la de menor capacidad. Pero si considerásemos esta densidad para todas las pistas sucedería que en las pistas más externas y de perímetro mayor los bits estarían muy separados produciéndose un cierto desaprovechamiento de la superficie magnética.

En conclusión el límite lo marca la pista interior. Si el radio interior es pequeño la superficie de almacenamiento es grande pero la densidad es pequeña, pero si el radio interior es grande la superficie de almacenamiento es pequeña pero la densidad es grande. Inmediatamente se plantea la siguiente cuestión: ¿cuál es el radio óptimo de la pista interna?, es decir, ¿cuál es el radio de la pista interna que producirá la máxima capacidad posible?

Para responder a esta cuestión se define el número de bits en cada pista $Nb_i = 2\pi R_i B_i$ y el número total de pistas $Nt = (R_e - R_i)T$, donde T es la densidad lineal de pistas ($T = \text{Pistas/cm}$). El número total de bits, N , será el producto del número de bits de la pista interior por el número de pistas en el margen escogido:

$$N = 2\pi B_i R_i T (R_e - R_i) = 2\pi B_i T (R_e R_i - R_i^2)$$

que es la ecuación de una parábola.

Derivando

$$\frac{dN}{dR_i} = 2\pi B_i T \frac{d}{dR_i} (R_e R_i - R_i^2) = 0 \Rightarrow R_e - 2R_i = 0 \Rightarrow R_i = \frac{R_e}{2}$$

Sustituyendo en la expresión de N :

$$N_{\text{máx}} = \pi B T \frac{R_e^2}{2} = \frac{\pi B T D^2}{8}$$

donde $D = 2R_e$. Por tanto aunque la superficie total del disco es $A = \frac{\pi D^2}{4} = \pi R_e^2$ y la

capacidad total sea $C_{\text{total}} = \frac{\pi B T D^2}{4} = \pi B T R_e^2$, incluso en el caso óptimo se tiene una eficiencia mucho menor:

$$\eta = \frac{N_{m\acute{a}x}}{C_{total}} = \frac{\pi B T D^2 / 8}{\pi B T D^2 / 4} = \frac{\pi B T R_e^2 / 2}{\pi B T R_e^2} = 50\%$$

2.11.1 Múltiples bandas

Dado que esta eficiencia es extremadamente baja es obvia la necesidad de obtener métodos que la mejoren. Una primera idea sería definir una sola pista con VLC (Velocidad Lineal Constante) en lugar de VAC (Velocidad Angular Constante). Esta es la técnica que se utiliza en los CD-DA y CD-ROM. Sin embargo para el caso de los dispositivos magnéticos presenta problemas en la velocidad de acceso debidos a dificultades de localización del sector buscado y problemas de inercia. Si la velocidad lineal se mantiene constante, al cambiar de una pista a otra, la velocidad angular debe variar, con lo que el disco deberá acelerar o frenar cada vez que hubiese un cambio de pista lo que provocaría unos accesos mucho más lentos, como sucede en los medios ópticos mencionados. Una segunda idea es realizar todas las pistas con la misma densidad pero tampoco es adecuada debido a que pistas contiguas no se diferenciarían en un sector completo sino en fragmentos de sector que no resulta eficiente.

La técnica que suele adoptarse para aumentar la eficiencia es conocida como “Banding”, dividir en bandas, y consiste en la división radial en zonas con distinta capacidad. De esta forma se consigue una gran superficie con alta densidad y es una solución de compromiso que aumenta la eficiencia del almacenamiento.

En esta técnica se consideran múltiples bandas tal que las distintas bandas mantienen la misma densidad lineal en la pista interior y todas las pistas de una misma banda tienen el mismo número de bits. Cada banda tiene un número de bits por pista creciente a medida que son más exteriores tal y como se muestra en la figura (2.12).

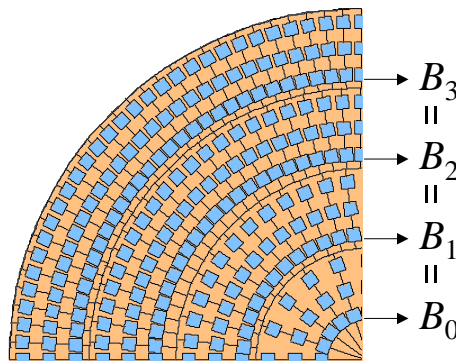


Fig. 2.12 Porción de un disco mostrando múltiples bandas

Para tratar el problema de manera general considérese que el área del disco está dividida en círculos concéntricos produciendo n bandas, donde n puede variar desde 1 hasta el número total de pistas.

Supóngase que todas las bandas tienen la misma anchura. Así el diámetro interno (d) o el radio interno (r) de una determinada banda (j -ésima) vendrá dado por:

$$d_j = d + j \left(\frac{D-d}{n} \right) \quad r_j = r + j \left(\frac{R-r}{n} \right)$$

y el radio externo de cada banda vendrá dado por el interno de la banda siguiente en sentido hacia el exterior:

$$R_j = r_{j+1}$$

Restando ambas expresiones:

$$R_j - r_j = \frac{R-r}{n}$$

Con esta configuración la capacidad de una banda, j , será:

$$N_j = 2\pi B T r_j (R_j - r_j)$$

y la capacidad total será la siguiente:

$$N_{Total} = \sum_{j=0}^{n-1} N_j = 2\pi B T \sum_{j=0}^{n-1} r_j (R_j - r_j) = \frac{2\pi B T (R-r)}{n} \sum_{j=0}^{n-1} r_j$$

Teniendo en cuenta que

$$\sum_{j=0}^{n-1} r_j = \sum_{j=0}^{n-1} nr + j \frac{R-r}{n} = nr + \frac{R-r}{n} \sum_{j=0}^{n-1} j = nr + \frac{R-r}{n} \frac{n(n-1)}{2}$$

se obtiene

$$N_{Total} = \frac{\pi B T}{n} [(n-1)R^2 + 2Rr - (n+1)r^2]$$

que es una parábola (siendo d la variable independiente). Para obtener el máximo número de bits se deriva la capacidad total y se iguala a cero

$$\frac{dN_{Total}}{dr} = \frac{\pi B T}{n} [2R - 2(n+1)r] = 0$$

despejando r al igual que se hizo con el disco de una única banda obtenemos:

$$r = \frac{R}{n+1}$$

y sustituyendo en la capacidad total se obtiene la capacidad óptima en función del número de bandas n :

$$N_{m\acute{a}x} = \pi B T R^2 \frac{n}{n+1}$$

y la eficiencia en el caso de múltiples bandas es: por tanto:

$$\eta = \frac{N_{m\acute{a}x}}{C_{total}} = \frac{n}{n+1}$$

Obsérvese que si n es grande la eficiencia se acerca al 100%. El máximo número de bandas se obtendrá cuando su número sea igual al número total de pistas. En este caso:

$$n = TR$$

La figura (2.13) muestra el número óptimo de bits en función del número de bandas, n . Como puede observarse el proceso es asintótico, es decir, la mayor ganancia en la capacidad se produce para un número dado de bandas después de las cuales el aumento en número de bandas produce un beneficio reducido. Por otro lado debe tenerse en cuenta que cuando el número de bandas crece, el diámetro de la pista interior disminuye, de modo que un aumento excesivo en el número de bandas se convierte en una situación impracticable.

En un disco multibanda, la velocidad de giro se mantiene constante para evitar los problemas de inercia mencionados anteriormente. Pero ahora, al contrario que en un disco con una sola banda, la velocidad de transferencia debe ajustarse puesto que cada banda requiere una velocidad de transferencia distinta lo que obliga a un diseño más elaborado del canal de lectura que engloba desde las cabezas al interfaz de conexión con el sistema principal. Esto es debido a que las celdas de bits de las bandas exteriores pasan más rápidamente por delante de la cabeza que las correspondientes a las pistas de las bandas interiores.

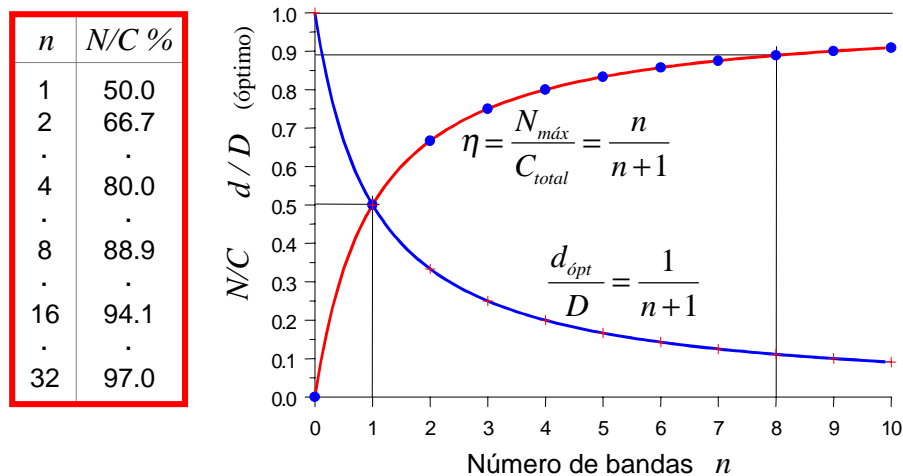


Fig. 2.13 Evolución de la mejora del aprovechamiento en función del número de bandas.

2.12 FORMATO DE GRABACIÓN

El término 'formato' describe el conjunto de reglas o procedimientos que deciden, cómo se representan y almacenan los datos en la superficie del disco. Podemos distinguir tres niveles.

En un primer nivel los bits de datos individuales se representan por cambios de magnetismo. El siguiente nivel, lo forman las cabezas que identifican los bloques representados en el disco con códigos usados para la detección y corrección de errores, y permiten etiquetar las distintas zonas del disco dividiendo en sectores, pistas, etc. El más alto de los niveles está relacionado con los archivos, los directorios y tablas de localización de ficheros, las cuales almacenan la localización de los bloques de ficheros de datos. El formato de este último nivel es realizado por el sistema

operativo y depende por tanto de las características que tenga este último. Nosotros veremos por el momento los dos primeros de estos niveles, es decir, los más bajos.

La información en la superficie del disco es grabada como una secuencia de inversiones de direcciones de magnetización en la superficie o flujos inversos; y es la posición de estos cambios de orientación magnética la que almacena la información. Existen varios modos de codificar la grabación de los datos, o lo que es lo mismo, existen múltiples formas de almacenar una misma secuencia de datos (bits) mediante patrones magnéticos. En la mayoría de estos modos podemos considerar la longitud de cada una de las pistas dividida en una secuencia de celdas de igual tamaño, cada una de las cuales almacena un bit de información. En este nivel no nos preocuparemos de si estos bits representan datos de usuario o alguna otra información como patrones de sincronización, encabezados de bloques, sector, pistas, códigos de detección de errores, etc. Este tipo de información que denominamos de control es utilizado por la unidad de disco para acceder a las distintas partes del mismo.

Ya se ha comentado que la celda de almacenamiento queda definida por el espacio necesario para que una transición pueda producir un pulso detectable. Si garantizamos que las transiciones tienen una separación mínima que reduzca el solapamiento, las celdas de almacenamiento podrán hacerse más pequeñas. Esto se consigue con una codificación adecuada. Por el contrario, si espaciamos demasiado las transiciones puede suceder que se pierda la sincronía de lectura. Los códigos RLL (Run Length Limited) acotan el espacio entre transiciones tanto por arriba (separación máxima) como por abajo (separación mínima).

La forma más obvia para almacenar un dato es magnetizar la superficie en una dirección para representar un '1', y en la dirección contraria para representar un '0'. Hay un flujo inverso para cada uno de los bits sólo si es diferente del bit precedente. A esto se llama código sin retorno a cero, o NRZ (Non Return to Zero). Su mayor inconveniente es que no define implícitamente la localización de cada celda de datos a lo largo de una cadena de bits con la misma polaridad magnética, lo cual es bastante frecuente. En estos casos aparece una amplia zona con una magnetización continua en una misma dirección. Esto significa que el número de bits en la cadena no puede ser determinado a partir de la información grabada, a menos que haya una señal de disparo externa o un reloj para definir la posición de cada una de las celdas bit. Esto estaba disponible en los primeros tambores. Algunas veces el reloj era grabado en una pista adicional y otras veces fue definido mecánicamente por unos dientes agregados a la rueda del tambor. Esto funcionó bien mientras la densidad de grabación fue baja; pero cuando la densidad de grabación fue incrementándose se hizo difícil evitar situaciones en las que el reloj y los canales de datos no tenían exactamente la misma longitud de paso debido a las finas imperfecciones eléctricas o mecánicas. El uso de los relojes separados de la pista es ahora inusual. Por lo tanto la grabación NRZ ha quedado en desuso para almacenamiento aunque se sigue utilizando en comunicaciones serie.

La variación de NRZ es la Inversión de No Retorno a Cero' (Non Return to Zero Invert) o NRZI. Aquí la variación del flujo ocurre cuando la celda representa un '1', pero no ocurre cuando esta representa un '0'. Esto se conoce como NRZI-Marca y alternativamente podemos definir NRZI-Espacio si las transiciones se producen en los ceros. Utilizado de forma aislada tiene el mismo problema que NRZ respecto a que no hay forma de contar cuantos ceros seguidos hay en una zona de magnetización constante. Obsérvese que las secuencias de unos no plantean problemas puesto que éstos introducen siempre una inversión de la magnetización. No obstante NRZI puede ser usado convenientemente cuando varios bits (normalmente un byte) son grabados en paralelo en pistas separadas, de tal forma que si añadimos un bit de paridad IMPAR en cada byte y lo almacenamos en una novena pista que se graba con una novena cabeza, entonces habrá por lo menos una pista con transición en cada localización de byte. Este modo de 'auto reloj', donde un grupo de pistas son tomadas de forma conjunta, se utiliza en las cintas magnéticas y algunos tambores de cabeza por pista, aunque de nuevo los problemas de sincronismo suelen aflorar con la

alta densidad de datos. En la mayoría de los discos cada pista es independiente, por lo que NRZI no es conveniente. En cualquier caso, esta solución no es aplicable a discos duros porque la alta densidad de grabación es muy superior a lo que permiten las tolerancias mecánicas entre las cabezas.

Las pistas individuales se vuelven 'auto relojes' si codificamos los datos de tal forma que haya al menos una inversión de flujo en un punto conocido de cada celda. El problema que plantea esta situación es que algunas celdas contendrán dos inversiones de flujo. Hay varios modos de grabación basados en este principio, pero todos ellos tienen una desventaja: que hay una mínima distancia entre flujos inversos en relación con el tamaño de la celda bit; es solo la mitad de como podría ser en el modo NRZ. Como las propiedades físicas del disco y la cabeza limitan el mínimo espaciado entre las inversiones de flujo este grupo de modos pueden grabar la mitad de bits que pueden grabar los modos NRZ. Por este motivo no se utilizan habitualmente, aunque si son utilizados en cintas.

Entre los modos de este último tipo, que garantizan una transición en una posición concreta de la celda, tenemos el de codificación en fase o modulación de fase (PE) y el de modulación en frecuencia (FM). El de modulación de fase consiste en garantizar que todas las celdas tienen una transición en el centro: ascendente si almacenan un uno o descendente si almacenan un cero o a la inversa. Si aparecen dos ceros seguidos o dos unos seguidos, se hace necesario añadir una transición extra al principio de la celda, para que las transiciones del centro de la celda se puedan llevar a cabo en el sentido correcto. La figura (2.14) muestra un ejemplo con distintos códigos.

Otra forma de codificación que también garantiza transiciones en todas las celdas es el de modulación en frecuencia (FM). Según este modo, todas las celdas tienen una transición al principio y añaden una segunda transición en el centro de la celda si almacenan un uno y no hacen nada si almacenan un cero o a la inversa. Este código se denomina de modulación o codificación en frecuencia porque la información de '1' o '0' se representa por la frecuencia de las transiciones. De esta forma las celdas con '1' tienen una frecuencia doble que las celdas con '0'.

Estos dos últimos métodos garantizan transiciones en todas las celdas, añadiendo una transición extra en algún punto conocido de la celda (el principio o el centro), y por lo tanto a la hora de la lectura se puede saber fácilmente cuantas celdas han pasado por delante de la cabeza en un intervalo de tiempo determinado. Esto puede realizarse con lógica secuencial sencilla, y el circuito que realiza esta función se denomina separador de datos. Su nombre proviene de la función que realiza: a la entrada se le proporciona la secuencia de impulsos magnéticos leídos por la cabeza debidamente acondicionados y tiene dos salidas, por una proporciona la secuencia de datos y por la otra la señal de reloj que ha extraído de la información de entrada. El problema que tienen estos códigos es que donde los códigos de tipo NRZ incluían una sola transición ahora se requieren dos, por lo que la capacidad se ve reducida a la mitad. Esto es consecuencia de que la distancia mínima entre transiciones viene fijada por el medio, la cabeza y otros parámetros de diseño y fabricación y es un límite físico que no se puede superar. En la figura (2.14) aparece un ejemplo con distintos códigos, donde se ve que para la misma información, los códigos PE y FM emplean el doble de transiciones que NRZ o NRZI.

El modo de grabación más utilizado en discos flexibles es el conocido como Modulación de Frecuencia Modificada (MFM) o Código Miller. Este código es una variación del código de frecuencia modulada. Como puede verse en la figura (2.14) el código FM incorpora una transición al principio de la celda, lo que provoca que las celdas que contienen un uno tengan dos transiciones o lo que es lo mismo la separación entre transiciones sea la mitad. El código MFM elimina esta transición al principio de la celda. Si se quedase así, tendríamos la misma situación que en NRZI donde no es posible saber cuantos ceros seguidos aparecen en una determinada secuencia. Para corregir esto, añade una transición al principio de las celdas de cero pero sólo si la celda anterior no incluyó transición en el centro. Es decir una celda de cero incluirá una transición

al principio sólo si va detrás de otra celda de cero. Por el contrario, las celdas de cero que van detrás de un uno no incluyen transición al principio, ya que si lo hicieran nuevamente tendríamos dos transiciones separadas por tan solo media celda. Como puede verse en la figura (2.14), la separación mínima entre transiciones vuelve a ser nuevamente de una celda completa al igual que en el caso de los códigos NRZ. Sin embargo y contrariamente a lo que sucedía con estos, nunca aparecerán largas secuencias de celdas sin transiciones. De esta forma, mediante una decodificación un poco más elaborada que la necesaria para FM se puede conseguir distinguir celdas individuales.

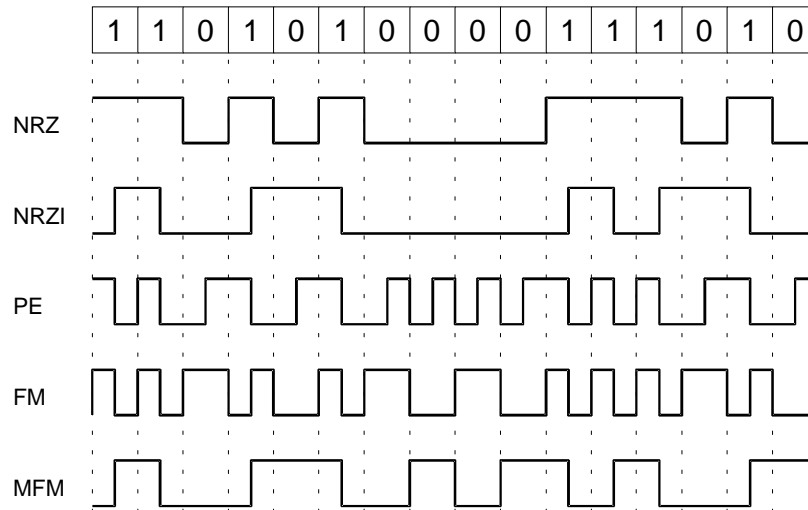


Fig. 2.14 Ejemplo de codificación de la información según distintos códigos

Podemos conseguir mayores densidades lineales con las mismas cabezas y medio si hacemos que el mínimo intervalo entre flujos inversos sea mayor que el ancho de una celda de bit. Haciendo esto permitimos que el máximo intervalo se vuelva mayor que el ancho de dos celdas. El tipo de codificación que usamos es llamado código limitado en longitud de recorrido o RLL (Run Length Limited coding). El nombre insinúa que nosotros diseñamos el código específicamente y emplazamos los límites inferiores y superiores de la longitud de cada carrera o secuencia de celdas de almacenamiento, las cuales no contienen transiciones de flujo. Como consecuencia la relación de los bits y celdas de almacenamiento se vuelve más compleja.

Para hacer uso de los códigos RLL hemos adoptado la técnica llamada grabación de código de grupo o GCR (Group Code Recording). Esto significa que en lugar de que cada uno de los bits de información corresponda a una celda de bit determinada, tomaremos un grupo de bits de datos juntos y los representaremos por un número de celdas de almacenamiento adyacentes. Es decir a cada grupo de la secuencia de datos de entrada, le asignamos un grupo de patrones de magnetización. Estos nuevos grupos asignados, tendrán algunas propiedades deseables que no tenían los grupos de datos originales. Una propiedad interesante es que las transiciones estarán separadas en un cierto número de celdas acotado tanto por arriba como por abajo. En los datos de partida no podemos imponer esto, puesto que los datos pueden contener cualquier secuencia arbitraria. Otra propiedad importante y complementaria de la anterior es que los nuevos grupos garantizan la presencia de alguna transición antes de un determinado espacio. Nuevamente es una circunstancia que tampoco podemos imponer a los datos de partida. Por este motivo grupos de datos son intercambiados por otros grupos con unas propiedades deseables. Mediante la primera de estas propiedades garantizamos que nunca aparecerán transiciones muy juntas con lo que podremos hacer las celdas de bit más pequeñas. Mediante la segunda, garantizamos que el circuito de lectura no perderá el sincronismo, ya que al limitar el número de celdas sin transición, se garantiza que aparecerá una transición en un determinado intervalo de tiempo.

Para que los grupos asignados, tengan esas propiedades, es necesario descartar aquellas combinaciones de bits que no las tengan. Como estamos descartando algunas combinaciones, los grupos asignados deberán tener una longitud mayor que los grupos de datos de partida.

El código de este tipo de más amplia difusión es el conocido como RLL-2,7 que garantiza que habrá un mínimo de 2 y un máximo de 7 celdas sin transición. En otras palabras, el mínimo espaciado entre inversiones es tres veces la longitud de la celda de almacenamiento y el máximo ocho veces. Este tipo de códigos son los que se emplean de forma prácticamente universal en los discos duros actuales aunque con distintos valores en sus parámetros.

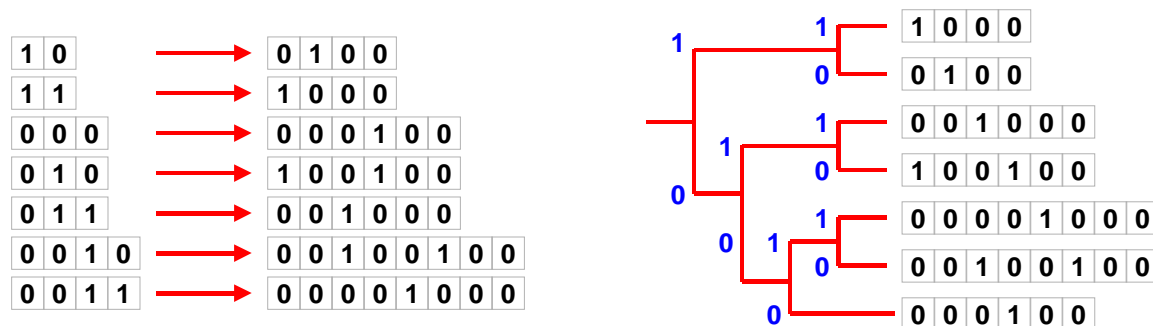


Fig. 2.15 Libro de códigos correspondiente al código RLL-2,7 junto con el árbol de codificación asociado, que ayuda en el diseño de un circuito codificador

El número de celdas de almacenamiento por grupo de datos es constante en algunas formas de código RLL y variable en otras como por ejemplo en el RLL-2,7 que como se ve en la figura (2.15) toma grupos de 2, 3 o 4 bits y les asigna grupos de 4, 6 y 8 celdas de código. Pero en cualquier caso el número de bits que representa será en promedio mayor. El libro de códigos no es único y puede elegirse de múltiples formas y en base a múltiples criterios. El presentado en la figura (2.15) es el propuesto por IBM para sus primeros discos 'Winchester' y está optimizado para reducir la propagación de un posible error de de/codificación. Sin embargo, la densidad que podemos conseguir en un disco particular está limitada por el espaciado entre flujos inversos, y en el código RLL este es varias veces el tamaño de la celda almacenada. Por ello, en realidad nosotros podemos almacenar más bits en la misma longitud de pista. En este código los bits de datos son grabados en cuatro celdas de almacenamiento. El código es más complejo puesto que tenemos que ver más de dos bits de datos a la vez, como tuvimos que ver dos bits juntos en MFM. El código MFM podemos considerarlo en realidad como un código RLL-1,3. Con la codificación RLL-2,7 podemos almacenar 1,5 bits de datos entre cada par de inversiones de flujo magnético. La ventaja se ve clara si tenemos en cuenta que para los códigos MFM teníamos un bit por transición y tan sólo medio bit para el PE o FM. El RLL-2,7 permite por tanto almacenar un 50% más de datos en el mismo espacio que el MFM. La densidad lineal de bit en los discos disponibles actualmente oscila entre 10000 y 40000 o más bits por pulgada y aún se incrementará más probablemente.

A la hora de escoger un determinado código de tipo GCR, hay numerosos factores a considerar. Por ejemplo la relación entre la frecuencia mínima y la máxima, que imponen restricciones al circuito codificador y decodificador respecto a su ancho de banda. También hay que tener en cuenta si los patrones magnéticos escogidos pueden introducir violaciones de código. Esto se produce cuando las condiciones de mínimo o máximo espaciado entre transiciones deja de cumplirse al poner un patrón a continuación de otro. Puede observarse en la figura (2.15) que cualquier concatenación de patrones sigue garantizando las propiedades del código. Esto no siempre es así, y por ejemplo, la codificación empleada en los discos ópticos (CD) no cumple esta propiedad y por lo tanto deben introducirse bits conectores.

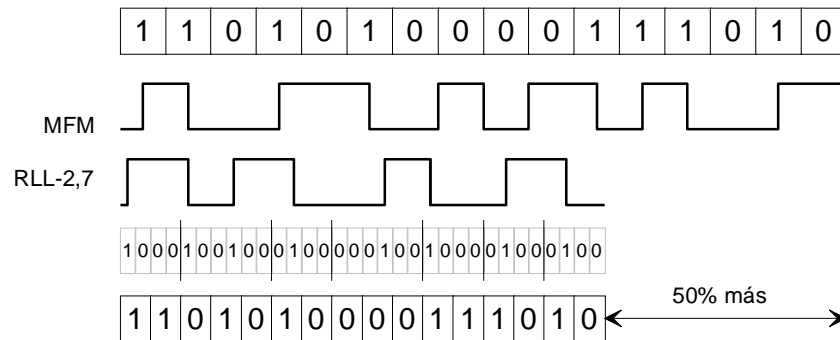


Fig. 2.16 Comparativa entre MFM y RLL-2,7

La figura (2.16) muestra claramente las ventajas de la codificación RLL-2,7. La misma secuencia de bits de datos ha sido almacenada en un espacio más reducido obteniéndose un aumento de capacidad de un 50% respecto a la codificación MFM. Esto se ha conseguido imponiendo una separación mínima entre las transiciones, lo que nos ha permitido juntarlas más. Como puede comprobarse en la figura, la distancia mínima entre transiciones es la misma que en el caso MFM. Aunque en la figura (2.16) no se aprecia, se puede comprobar, viendo los códigos asignados que aparecen en la figura (2.15), que se garantiza la existencia de al menos una transición en cada grupo, puesto que todos los grupos asignados incluyen al menos una. También se puede comprobar que la separación mínima entre dos transiciones es de dos celdas y la separación máxima de celdas sin transición es de siete. El precio que hay que pagar por este incremento de capacidad es la complejidad. La codificación y decodificación es considerablemente más compleja. Por otra parte este código no puede considerarse como un código con autorreloj y por lo tanto no puede emplearse un circuito separador de datos para obtener la secuencia de datos y la señal de reloj. Para obtener la señal de reloj se precisa de un circuito especial que sea capaz de generar la señal de reloj en sincronía con las transiciones y mantener la frecuencia sin derivas en los intervalos entre transiciones. Un circuito capaz de realizar estas acciones se conoce como PLL (Phase Latched Loop) bucle de enganche de fase o circuito de fase sincronizada, que permite mantener una oscilación en fase con otra. El funcionamiento de un circuito de este tipo puede encontrarse en cualquier texto de electrónica.

2.13 SERVOPISTAS

Como ya se comentó antes, muchos discos usan servo control para situar la cabeza y asegurar que la cabeza está posicionada en la pista requerida. Esto se vuelve más necesario cuando el espaciado entre las pistas decrece, y no es tan fácil reducir la tolerancia de los mecanismos en la misma proporción. Los servomecanismos necesitan una señal de entrada para determinar cuanto debe desplazarse el brazo para situar la cabeza sobre la pista de datos. Hay dos métodos por los que usualmente se proporciona al servo esta información. Uno de ellos consiste en reservar una superficie del disco para esta información y esto se hace incluyendo una servopista separada en cada cilindro. A este método se le llama método de la superficie servo-dedicada. Con esta disposición se consigue que el canal de lectura y la cabeza del servo estén completamente separados de la cabeza de datos y canal de lectura/escritura, obteniéndose un diseño más simple del mismo, pues no precisa de un separador de datos. Sin embargo, significa que una superficie que podría ser utilizada completamente para datos no lo sea, lo cual es un inconveniente, en especial en los dispositivos que utilizan un número reducido de platos. En un disco de un solo plato se pierde un 50% de la superficie total que podría dedicarse a datos ya que la superficie servodedicada corresponde a una de las dos caras del disco. En discos con múltiples platos, sin embargo, este inconveniente no es tan fuerte puesto que existen muchas más caras (dos por cada plato) pero sólo una cara de un solo plato se reserva como superficie servodedicada por lo que el porcentaje de superficie que puede utilizarse para datos es considerablemente mayor que en el caso

de un disco simple y será mayor cuanto mayor sea el número de platos. Este método supone también que todas las cabezas del cilindro permanecen muy cerca de la misma posición relativa de uno a otro, y esta presunción puede tener poco éxito en los dispositivos con muchos platos y espaciado de pista reducido (principalmente controladores de disco con medios intercambiables). La alternativa es incluir la información de servo en cada una de las pistas de datos y leerlo con la misma cabeza que se lea o escriba en la pista. A estas técnicas se les llama servotécnicas embebidas ('embed servo techniques').

Existen varias maneras de proporcionar la información necesaria al servo. La más simple, a menudo llamada servocuña, consiste en que la información aparece sólo en un punto de la pista (en el área de índice, entre el final del último bloque en la pista y el comienzo del primero). El servo sistema usa esto para determinar la posición correcta, y entonces retiene esta información para el resto de la revolución. Esto es bastante bueno para modelar el espaciado de pistas, pero incrementa los tiempos de acceso porque la cabeza debe esperar a que el servo vuelva a encontrarse en la pista, y entonces esperar de nuevo al sector requerido con lo que el tiempo de latencia se duplica. Esto se puede reducir si al comienzo de todos los sectores se introduce una servocuña, pero en este caso se reduce la capacidad nominal del dispositivo en beneficio de una mayor velocidad.

2.14 FORMATO DE LA PISTA

Hemos descrito el menor nivel de definición del formato, el cual determina cómo cada uno de los bits de información está representado con un patrón de inversiones de flujo magnético a lo largo de la pista. A este nivel no hay distinción entre bits que representan datos del usuario y aquellos que son agregados por la unidad de disco y su controlador (encabezamientos de sectores y caracteres de sincronismo por ejemplo). Podemos considerar ahora el siguiente nivel de formateo, donde la pista es dividida en un número de sectores o bloques separados.

Hemos visto que los discos y tambores son dispositivos de transferencia por bloques; transferencia desde y hacia la unidad básica de un bloque de datos a la vez, a diferencia de las impresoras, por ejemplo, que manejan la información carácter a carácter. En la mayoría de los discos y tambores el tamaño del bloque es constante y se determina cuando se diseña la unidad de disco. Pocos fabricantes, usan formatos con bloques de longitud variable, y cada bloque incluye una zona de pista para definir su longitud. Nosotros supondremos que la longitud del bloque es fija en la descripción que viene a continuación.

Varios bloques son escritos, uno detrás de otro, en cada una de las pistas del disco; por lo tanto, cada uno ocupa un sector de la pista. En el contexto del disco, se usará el término sector en lugar de bloque. Por razones que explicaremos luego, el número de sectores por pista es habitualmente un número primo; en discos duros de moderada capacidad es a menudo 17. La capacidad del sector puede variar, pero en pequeños discos duros es de 512 bytes. La velocidad de rotación habitual para tales discos es 3600 r.p.m., lo cual da un promedio de datos de 520 Kbytes o 4.16 Megabits por segundo. Como cada pista lleva alguna información adicional junto a los datos de usuario, este valor se incrementa hasta unos 5 Megabits por segundo. Algunas de las interfaces más usadas en estos discos especifican el rango de bits, y de ahí (a menos que la velocidad de rotación cambie) el número de bits por pista. Por lo tanto, la capacidad del dispositivo puede variar sólo cambiando el número de pistas por superficie del disco y el número de superficies usadas.

La posición de cada sector está definida por el encabezamiento, que es escrito antes de que el disco sea usado por primera vez. El proceso se denomina formateado de bajo nivel, que puede ser reescrito después si es necesario (en el caso de dispositivos que utilicen servo técnicas embebidas; esto requerirá un equipamiento especial, aunque también puede realizarse con un programa de utilidad). Pero este formateo físico o de bajo nivel no es normalmente alterado y cada

uno de los sectores permanece en la misma posición a través de toda la vida del disco. Este formateo a bajo nivel se reescribe a veces porque los niveles de señal tienden a fallar ligeramente a lo largo del tiempo y esto puede reducir la fiabilidad. Sin embargo, la posición del sector no se altera cuando se realiza un formateo a alto nivel y que discutiremos más adelante.

Algunos discos modernos utilizan geometrías complejas en las que el número de sectores por pista no es único. Piénsese por ejemplo en los discos que emplean 'banding'. En estos casos debe evitarse el formateo a bajo nivel o debe hacerse con precaución, ya que unos parámetros incorrectos podrían destruir la estructura o geometría original, a menos que se disponga de un programa de utilidad que contemple las características específicas del modelo de disco. Sin embargo, la velocidad con la que los fabricantes modifican los diseños y la baja disponibilidad a proporcionar esta información hacen imposible realizar un formateo adecuado.

Cada una de las pistas comienza con una marca de orden. Hay solo una marca de orden por cada pista y las marcas para todas las pistas están en la misma línea radial. Estas marcas pueden ser escritas en cada pista o donde hay dedicada una servo superficie. Los discos flexibles tienen insertadas las marcas de orden físicamente. Hay alguna marca física de alguna clase, normalmente un agujero en el disco.

El primer sector comienza poco después de la marca de orden y los restantes sectores están espaciados igualmente a lo largo de la pista. Cada sector está dividido en cabeceras y bloques de datos. Las cabeceras son escritas como parte del formato de bajo nivel, y de ahí en adelante tratados como permanentes y los bloques de datos, los cuales son escritos durante el uso normal del disco pueden ser reescritos cuando se desee. Cabecera y datos están precedidos por unos pocos caracteres de sincronismo y seguidos por dos o más bytes de chequeo. Los bytes de chequeo son usados en el manejo de errores (que describiremos brevemente).

Entre cabeceras y datos, entre sectores y entre marcas de orden y sectores adyacentes están los 'gaps', cuya longitud puede variar dentro de ciertos límites. Estos gaps separan secciones escritas en la pista en diferentes operaciones. Ellos son necesarios porque la velocidad del disco puede no ser la misma en cada ocasión, y por lo tanto, la longitud de cada sección puede variar ligeramente. Su longitud puede no ser un múltiplo exacto del espaciamiento de caracteres, ya que son necesarios los caracteres de sincronismo. Cada gap es equivalente a unos pocos caracteres, excepto el gap 4, que es considerablemente mayor porque rellena el valor completo del sector.

El contenido de la cabecera del sector varía ligeramente de un diseño a otro, pero siempre incluirá el número de cabeza y número de cilindro (y por tanto de pista), y también el número de sector dentro de la pista; estos también estarán espaciados por un bit de estado para identificar los sectores defectuosos, si es necesario. En aquellos formatos donde la longitud del sector es variable, la cabecera contendrá un espacio para el número de bytes de datos en el bloque. Cuando el disco es formateado (sólo a bajo nivel, excepto los discos flexibles) frecuentemente, el área de datos es llenada con caracteres falsos o blancos.

Cuando el formateo a bajo nivel ha sido realizado (Fig. 2.17), el disco es leído para chequear que todas las cabeceras y caracteres de datos puedan ser leídos correctamente. Las condiciones de lectura son más estrictas que en su uso normal. Por lo tanto, cualquier sector que se encuentra satisfactorio en este momento, es improbable que dé problemas en el uso posterior. Si el sector es encontrado no satisfactorio, su cabecera es reescrita con un bit de estado apropiado. El controlador del disco reconocerá entonces que el sector está inutilizado cuando él lea la cabecera y lo sustituirá en otro sector reservado para este propósito. Los fabricantes de discos usualmente suministran una lista con sectores en mal estado en la unidad de disco, y puede ser introducida en la controladora; los programas que realizan un formateo a bajo nivel, o simplemente un test del disco, producirán una lista actualizada de sectores defectuosos. Interfaces inteligentes, tales como el SCSI, pueden ocultar la existencia de sectores en mal estado al usuario.

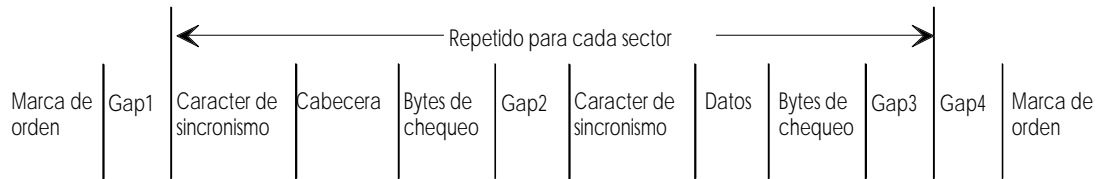


Fig. 2.17 Formateo a bajo nivel de una pista

2.15 ENTRELAZADO

Algunas controladoras de disco, o sus procesadores principales, son incapaces de leer o escribir dos sectores en una rápida sucesión. Si el procesador principal quiere leer dos sectores sucesivos, el segundo será enfrentado a la cabeza antes de que la controladora esté preparada para ello, y por lo tanto no puede ser leído hasta que el sector se enfrente a la cabeza en la siguiente revolución. Esto reduce drásticamente la velocidad de transferencia del disco. Por ejemplo, si hay 17 sectores por pista, la velocidad de transferencia de datos será reducido por un factor de 18.

La solución a este problema se resuelve en el formateo a bajo nivel, donde los sectores son entrelazados. En lugar de que los sectores estén numerados consecutivamente desde la 'marca de orden', el sector 2 es el tercer sector a lo largo de la pista, el sector 3 es el quinto y así sucesivamente (ver figura 2.18). Esto da un factor de entrelazado de 2, y la velocidad de transferencia de datos es ahora reducida por un factor de dos respecto del valor nominal. Nosotros podríamos usar un mayor factor de entrelazado para dar más tiempo a que el controlador se prepare. Por ejemplo si el entrelazado es 3, los sectores serán renumerados: 1-7-13-2-8-14-3-9-15-4-10-16-5-11-17-6-12. Esto permite que la CPU almacene los datos para el sector 1 mientras que los sectores 7 y 13 están enfrentados con la cabeza y continuar con el 2 cuando ya está preparado. En una revolución se habrán leído y almacenado aproximadamente 6 sectores por la CPU; y en tres revoluciones se habrán leído y almacenado todos los sectores. Si el número de sectores por pista es un número primo, podemos usar un factor de entrelazado más pequeño que el número de sectores sin llegar al mismo sector físico antes de que hayamos localizado todos los sectores.

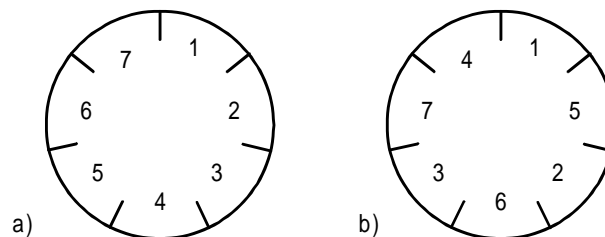


Fig. 2.18 Entrelazado de sectores

Para todas las configuraciones, hay un factor óptimo de entrelazado. Los valores más pequeños reducirán drásticamente la velocidad de transferencia, mientras que los valores mayores provocarán transferencias más lentas ya que contribuyen a aumentar el tiempo de latencia efectivo. De este modo, el factor óptimo puede depender de la configuración del procesador principal, tanto como de la controladora del disco; un entrelazado inadecuado del disco duro al reformatearlo puede producir unas prestaciones inferiores a las óptimas. Algunos discos incluyen una memoria FIFO capaz de almacenar varios sectores, normalmente una o dos pistas completas, de tal forma que puede trabajar sin entrelazado (entrelazado de 1) aunque estén conectados a interfaces lentas.

2.16 TRATAMIENTO DE ERRORES

La integridad de los datos almacenados es de gran importancia para el usuario. Un simple dígito erróneo podría tener resultados catastróficos en ordenadores usados en defensa o en empresas financieras. Sin embargo en los sistemas que manejan datos muy redundantemente como textos y especialmente sonido o imágenes codificadas como mapas de bits, el resultado de un dígito erróneo sería menos crítico e incluso podría pasar desapercibido. Desafortunadamente, ni los medios de almacenamiento ni los dispositivos son perfectos. Las unidades de disco y sus controladoras pueden ser diseñados para detectar y reconvertir la mayoría de los errores de datos, pero en la mayoría de los casos, esto incrementa excesivamente el coste. El diseñador del sistema debe llegar a un compromiso entre la integridad de los datos y el coste del sistema. Para la mayoría de los sistemas, la cantidad gastada en el dispositivo de almacenamiento no conseguirá la integridad requerida, por lo que el software debe proporcionar el mayor nivel de detección de errores. Existen varios métodos para hacer esto -el uso de 'checksums' es uno de los más conocidos, pero el más habitual es el código de redundancia cíclica o CRC que se describirá más adelante.

Lo más importante y esencial en el manejo de errores es conocer la magnitud del problema. La desviación de la unidad de disco (u otro dispositivo de almacenamiento) de la perfección es medida en términos de la magnitud del error, que es la razón entre el número de bits leídos o escritos por el dispositivo y el número de errores ocurridos a lo largo de estos bits. La magnitud es normalmente expresada como 1 en 10^n significando eso que no más de un error ocurrirá para 10^n bits procesados. En este contexto, 'un error' no significa necesariamente un error en un bit sino que se toma usualmente para cubrir un grupo de bits adyacentes afectados, por ejemplo, un simple error en el medio o más generalmente, algún grupo de errores que puede ser reconvertido con una operación simple -tal como la relectura de un sector del disco-. La definición exacta del error rara vez se da en las especificaciones; por lo que las tasas de errores son una anotación poco precisa.

Existen varias formas de evaluar los errores, pero de cara al usuario, los más importantes son dos: 'undetected error rate' que mide errores no detectados por la unidad de disco o su controladora (estos errores pueden ser detectados por supuesto en cualquier otro sitio del sistema), y los 'irrecoverable o permanent o uncorrected error rate' (usualmente llamados 'hard error rate'), que mide errores que el dispositivo detecta pero no puede corregir. Otros tipos de errores que el usuario puede conocer son los 'recoverable or transient or corrected error rate' (o 'soft error rate'), refiriéndose a los errores que el propio dispositivo puede corregir, con o sin ayuda del sistema operativo, y el 'seek error rate'. El último se expresa como 1 en 10^n búsquedas; se refiere a las ocasiones en las que el dispositivo busca una pista (por ejemplo, mueve la cabeza a la posición de lectura de la pista), y durante la lectura de las cabeceras encuentra que está en una pista equivocada. Los 'soft errors' y los 'seek errors' no conciernen realmente al usuario, puesto que se corrigen automáticamente, aunque si hay muchos errores, repercutirán en las prestaciones del sistema. Sofisticados sistemas almacenan el número de estos errores como una medida del buen funcionamiento del dispositivo.

El 'hard error rate' se da normalmente en las especificaciones del dispositivo, y para discos magnéticos (discos duros y discos flexibles) es usualmente de 1 en 10^{12} , aunque algunos fabricantes den un valor de 1 en 10^{13} para sus discos duros. Para discos rápidos que transfieren datos a razón de un megabyte por segundo, 1 en 10^{13} representa un error cada 350 horas de transferencia de datos. El 'undetected error rate' se especifica muy raramente, particularmente por su dificultad a ser medido o predicho, pero se asume usualmente como 100 veces mejor que el valor de 'hard error'. El 'soft error rate' es usualmente fijado a 1 en 10^9 para los discos flexibles y 1 en 10^{10} (y ocasionalmente 1 en 10^{11}) para los discos duros. Los 'seek errors' usualmente rondan el 1 en 10^6 búsquedas, y ocasionalmente el 1 en 10^7 búsquedas.

Hemos estado considerando implícitamente los errores de lectura, suponiendo que los datos se escribieron correctamente en el disco en el primer sitio y que no ocurre ningún fallo almacenando o escribiendo. Se producen también fallos de escritura, donde el dispositivo falla almacenando los datos incorrectamente. Sin embargo, todas las unidades de disco tienen previsto un chequeo por si se producen errores durante la escritura, verificando el bloque o pista en la próxima revolución del disco. Por lo tanto, los errores de escritura no perjudican la integridad del dato, aunque pueden disminuir las prestaciones. Un gran número de errores de escritura significa normalmente que el medio está llegando al final de su vida útil. El rango de errores de escritura se especifica algunas veces, pero no muy a menudo.

La evaluación de errores se realiza en dos etapas. Primeramente el error debe ser detectado, y sólo entonces puede ser posible arreglarlo. Este proceso se entiende mejor si consideramos por separado estas dos etapas.

La detección del error depende de la existencia de algún grado de redundancia en el dato según se graba en el disco; en otras palabras, almacenamos un mayor número de bits respecto al mínimo necesario para almacenar un dato. Hay muchas formas de hacer esto. La más simple y quizá la más vieja de éstas es la paridad impar. Esto conlleva añadir un bit extra a cada una de las unidades de datos (usualmente a cada byte). El bit es escogido de tal forma que el número de veces que aparece '1' sea siempre impar. Cuando leemos el dato, chequeamos cada byte y si hay un número par de '1' sabremos que ha ocurrido un error. No podemos saber qué bits dan error, ni tampoco podemos estar seguros de que la paridad nos detecte todos los errores en el byte, por lo que la paridad impar proporciona sólo una detección simple de errores (SEC). La paridad es muy útil donde el byte es grabado como una unidad discreta o estructura con cada uno de los bits en pistas separadas, como es el caso de las cintas magnéticas; siendo menos útil en los discos, donde cada dato se graba como un conjunto de bits en serie en cada pista. En algunos de los primeros tambores, que grababan los datos en conjuntos de bits en paralelo, también se utilizó la paridad.

Como hemos visto, los datos son grabados en el disco en bloques de un Kbyte más o menos, con cada byte separado serialmente a lo largo de la pista. El bloque es por tanto grabado como una cadena simple de bits. Aunque podríamos chequear la paridad de la cadena como un todo, podría ser de escaso valor, porque el espaciado entre bits es pequeño, y por lo tanto la probabilidad de que el fallo afecte a más de un bit es alto. Entonces podríamos aplicar una paridad de bits adicional basada en subsecuencias de bits de datos. Por ejemplo, la segunda paridad podría ser calculada usando solo los bits primero, tercero, quinto, y así a lo largo de la cadena de datos. Podemos continuar este proceso con los bits de datos seleccionados por otros caminos, de este modo los bits de paridad extra disminuyen el número de errores en el bloque que estamos chequeando. Sin embargo, esto es un proceso caro, debido al número de cálculos separados que estamos obligados a hacer. Podemos conseguir un efecto similar haciendo un cálculo más simple, y éste es el método de detección de errores que más ampliamente se utiliza en las controladoras de disco.

Este método se conoce como chequeo de redundancia cíclica o CRC (Cyclic Redundancy Check). Ahora, al número de bits de paridad, agregamos un número binario llamado CRC al final de cada bloque de datos; un CRC de 2 bytes es suficiente para todas las longitudes de bloque normales. El CRC se calcula en principio como una función de la cadena de datos del bloque, visto como un número binario simple y también a partir de una potente serie conocida como 'polinomio generador'. Pueden utilizarse muchos polinomios; uno de los más populares es escrito como $x^{16} + x^{12} + x^5 + 1$; la función es tal que si aplicamos la misma función a la cadena de bits (incluyendo el CRC) leída del disco, el resultado debería ser cero. Si no lo es, entenderemos que existe un error en el bloque; y si es cero, no podemos estar absolutamente seguros de que no haya errores, puesto que existe aún la posibilidad de que el conjunto de errores ocurridos en total, tenga

efectos tales que unos cancelen a otros. Sin embargo, puede calcularse la probabilidad de que tal conjunto de errores ocurra, y es muy baja si el polinomio generador es adecuado.

La operación a realizar es en principio una división y podría ser implementada de esa forma, pero de hecho, esto puede conseguirse mucho más fácilmente con un hardware dedicado en forma de registro de desplazamiento. El registro sólo necesita tantas etapas como bits CRC, usualmente 16. Cuando escribimos, la cadena de bits que forma el bloque se introduce dentro del registro al mismo tiempo que se escribe en el disco. Cuando toda la cadena ha sido introducida, el contenido del registro es utilizado como CRC. Durante la lectura, se usa el mismo hardware; el dato y el CRC agregado se pasan a través del registro y su contenido final debe ser comparado con cero.

En el caso de los errores de escritura de hecho, sólo podemos saber si los datos se han escrito correctamente leyéndolos, y por supuesto, el error puede ocurrir cuando hacemos la lectura de chequeo de los datos (en la lectura de chequeo no comparamos los datos leídos con los datos originales; usamos simplemente el CRC para comprobar si existen o no errores en los datos leídos del disco). Por lo tanto, la detección de errores de escritura usa exactamente el mismo proceso en la detección de errores. Sólo podemos distinguir entre los dos primeros suponiendo que el error se produce en la lectura; si falla en respuesta a las técnicas de recuperación que usamos para los errores de lectura, entonces comenzamos a tratar el error como de escritura.

Hemos visto cómo detectar errores; ahora vamos a considerar cómo podemos tratarlos en el proceso conocido como recuperación del error. Hay básicamente dos métodos. El primero consiste simplemente en tratar de leer o escribir el bloque completo y lo llamamos reintento ('retry'). El otro método usa un código redundante de datos para identificar en qué bit particular ocurre el fallo y entonces corregirlo: esto es la corrección del error. Todas las controladoras de disco magnético utilizan el primer método, y muchas también usan el segundo variando el nivel de corrección. Aunque la corrección de errores esté disponible, es más rápida y eficaz la relectura y por lo tanto, se intenta primero. Las unidades de discos con interfaces inteligentes, realizarán la relectura de forma automática y el procesador central no será consciente de que ocurre el error, excepto en una cierta oscilación en el flujo de datos. Otras unidades de disco, necesitarán de la ayuda del sistema operativo. Si el dispositivo detecta un error en el bloque que ha leído, no lo pasa al procesador principal, pero sí señala el error, y el procesador principal da el comando necesario para que se realice una nueva lectura del bloque. Esto se repite si es necesario una serie de veces, a menudo hasta 10 intentos. En el caso de que el bloque aún no se haya leído correctamente el procesador central activa un 'hard-error'.

La corrección de errores, en contraste con 'retry', no involucra al sistema operativo y depende de si existe bastante redundancia en el código de datos, para permitir al dispositivo encontrar qué bit particular es el erróneo, y entonces, corregirlo. Cualquier esquema de corrección de errores particular puede corregir sólo cierto número de bits; si hay más errores en este bloque, la unidad tiene que recurrir a 'retry'. En definitiva, el número de bits que se pueden detectar es proporcional al grado de redundancia en la cadena de datos, aunque algunos métodos de codificación son mejores que otros a este respecto.

El CRC discutido antes proporciona un cierto grado en la corrección de errores si el contenido del registro de desplazamiento no es cero después de la lectura del bloque y su CRC. En ese caso se demuestra la existencia de un error; y el contenido es, de hecho, un indicativo de la dirección del bit erróneo, con tal de que sea uno sólo. Sin embargo, sólo podemos aprovechar esta información si podemos demostrar que sólo ha habido un error. Para ello realizamos un segundo CRC (de hecho, uno de ellos se describe como Error Correcting Code o ECC), calculado de otra manera. El ECC se usa para corregir el bloque en el que se supone que sólo ha ocurrido un error; el CRC chequea si el bloque corregido es realmente correcto.

La mayoría de las unidades de disco sólo llegan hasta aquí, pero es posible usar otros métodos de codificación más elaborados que permiten corregir errores mucho más extensos. Esto involucra la recodificación de los bytes de datos, o cambiar el orden de los bits o bytes (códigos de entrelazado cruzado). Esta última técnica es empleada por ejemplo en los CD's de música.

2.17 CÁLCULO DEL CRC

En el argot CRC, los mensajes se consideran como largos polinomios dentro de los cuales cada bit 0 ó 1 se expresa como el coeficiente de un término de los mismos. El exponente de cada término se obtiene de la posición ordinal del bit dentro del mensaje.

Ej.: 01011010 = $0x^7 + 1x^6 + 0x^5 + 1x^4 + 1x^3 + 0x^2 + 1x^1 + 0x^0$
 los términos cero se omiten: $x^6 + x^4 + x^3 + x^1$

El polinomio mensaje se divide por otro llamado polinomio generador, produciendo un cociente y un resto. La división se realiza con aritmética módulo 2: en lugar de una resta ordinaria se realiza una operación XOR sin tener en cuenta los arrastres. El resto de la división no se transforma en el CRC hasta que se 'limpia' añadiéndose un bit 0 al polinomio mensaje por cada término del resto. Por lo tanto, en un CRC de 16 bits, el dividendo se rellena con 16 bits 0. El valor de comparación CRC es precisamente el resto de esta división módulo 2 del polinomio mensaje relleno adecuadamente; el cociente, por su parte, se descarga.

Los dos polinomios generadores de 16 bits que se usan más frecuentemente son:

- El polinomio CCITT: $x^{16} + x^{12} + x^5 + 1$
- El polinomio CRC-16: $x^{16} + x^{15} + x^2 + 1$

2.17.1 División polinómica por hardware

Se realiza usando circuitos biestables ('flip-flop') y puertas OR-Exclusivas. En la figura (2.19) se muestra el circuito hardware clásico que realiza esta operación empleando el polinomio CCITT. En este caso, el bit de datos más significativo se introduce en el bit bajo del registro resto; este registro se desplaza a la izquierda en cada etapa. Después de la división, el valor del registro corresponde exactamente al resto obtenido en la división larga. Este circuito divide un polinomio mensaje de cualquier longitud por el polinomio de 17 bits especificado de la siguiente manera:

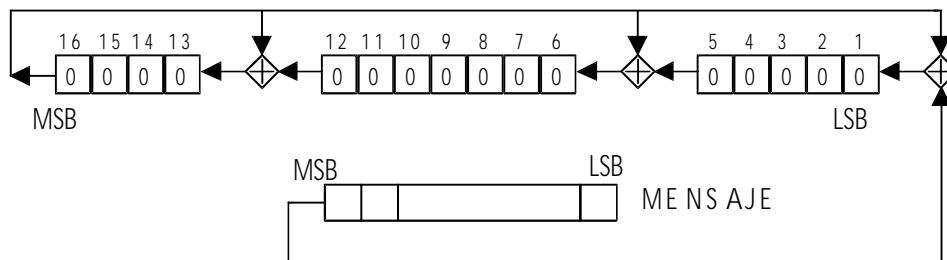


Fig. 2.19 Hardware clásico que realiza la operación de división con el polinomio CCITT. MSB y LSB indican el bit más significativo y el menos significativo del mensaje y del registro respectivamente.

- 1.- Se introduce el bit de datos de más peso en el registro resto.
- 2.- Se desplaza el bit en el registro desde el más bajo. El bit más alto del resto se desplaza hacia la izquierda. En realidad, estos bits altos constituyen el cociente, que carece de interés para nosotros.

- 3.- Si el bit que salió del resto era TRUE, se sustrae el polinomio divisor completo (XOR) del resto.
- 4.- El byte de datos se desplaza a la izquierda un lugar. El bit que se expulsa no se utiliza.
- 5.- Se repiten los pasos 1-4 hasta haber aplicado todos los bits al registro resto.
- 6.- El registro contiene ahora el resto de la división polinómica.

2.17.2 Aritmética en módulo 2

En primer lugar, hemos de establecer que el punto de arranque para diseñar cualquier sistema CRC es el número de bits deseado en el valor de comprobación (normalmente 16).

Vamos a ver en un ejemplo el funcionamiento del CRC. Para ello, supongamos un mensaje de cuatro bytes (CFYU):

C=11000011 F=01100110 Y=11111001 U=01010101

Supongamos también que queremos transmitir un valor de comprobación de 16 bits. Consideremos el mensaje como un sólo número:

$$11000011011001101111100101010101 = 3.278.305.621$$

El número de comprobación se obtiene dividiendo este número por otro que en este ejemplo es 525:

$$3.278.305.621 / 525 = 624.439 \quad \text{y cuyo resto vale: } 346$$

El problema que surge ahora es el siguiente: este cociente de 23 bits es un valor de comprobación perfectamente aceptable, pero tendremos que truncarlo a 16 bits para poder enviarlo, por lo que su precisión queda en entredicho. Ya que no podemos asegurar la longitud del cociente, podemos probar con el resto ya que con este sí podemos asegurar cuál será su longitud máxima: si el divisor es de 17 bits, el resto no tendrá más de 16 bits:

$$3.278.305.621 / 65.540 = 50019 \quad \text{con resto} = 60361$$

que puede expresarse en 16 bits: 60361=EBC9=1110 1011 1100 1001

2.17.3 División larga en módulo 2

El resto del valor de comprobación no se obtiene en aritmética binaria usual, sino en módulo 2, lo cual simplifica considerablemente el hardware, ya que esta aritmética carece de acarreo y se realiza de igual manera que la operación OR-Exclusiva. Así tenemos que la suma es igual a la resta.

La única diferencia mecánica entre la división módulo 2 y la división binaria ordinaria es que los resultados intermedios se obtienen mediante operaciones OR-Exclusivo en lugar de sustracciones. Se trata de forzar una división binaria en la que el bit situado a la izquierda en el resto anterior se hace 0. A continuación se puede observar cómo los bits de mensaje van añadiéndose uno a uno por la derecha, obteniéndose resultados intermedios. Si el bit de mayor orden del resto intermedio es 1, se envía 1 al cociente y se resta el divisor (XOR) del resto. Por el contrario, si el primer bit es 0, se envía el 0 y se le resta 0 (de 16 bits).

11000011011001101111100101010101	10001000000100001
<u>10001000000100001</u>	1100111110000111
x10010110111011001	cociente
<u>1000100000010000</u>	
x00111101111110001	
<u>0000000000000000</u>	
x01111011111100011	
<u>0000000000000000</u>	
x11110111111000111	
<u>10001000000100001</u>	
x11111111111001100	
<u>10001000000100001</u>	
x11101111111011010	
<u>10001000000100001</u>	
x1100111111110111	
<u>10001000000100001</u>	
x1000111110101100	
<u>10001000000100001</u>	
x00001111100011011	
<u>0000000000000000</u>	
x00011111000110110	
<u>0000000000000000</u>	
x00111110001101101	
<u>0000000000000000</u>	
x01111100011011010	
<u>0000000000000000</u>	
x11111000110110101	
<u>10001000000100001</u>	
x11100001100101000	
<u>10001000000100001</u>	
x11010011000010011	
<u>10001000000100001</u>	
x1011011000110010	← Resto final

La x no indica multiplicación sino cada uno de los ceros que se cancelan en cada etapa. (Salen del registro de desplazamiento)

Fig. 2.20 Ejemplo de división módulo 2 empleando el mensaje de datos como dividendo y el polinomio CCITT como divisor

2.18 FORMATO DE ALTO NIVEL

El tercer y mayor nivel de formateo del disco (el formato lógico), está relacionado con la utilización de los sectores para elementos específicos de los datos. En este nivel, el formato está determinado por el sistema operativo (S.O.) del procesador principal, más que por el controlador y la unidad de disco. Lo habitual es reservar los primeros sectores de cada disco para usarlos por el sistema operativo. De estos sectores, uno o más se usarán como directorio, y también para las tablas y estructuras que permiten la ubicación de ficheros.

El usuario esperará que este sistema maneje los ficheros de datos, que pueden ser de cualquier longitud, mediante nombres identificadores (aunque cuando los escribimos en el disco, ocuparán sectores completos), y parte de la tarea del S.O. es decidir dónde guardar este archivo en el disco. El directorio es una lista que contiene el nombre y longitud de cada uno de los ficheros que se han escrito en el disco, y también la dirección de su primer bloque (número de cabeza, número de cilindro, y número de sector). Inicialmente el primer fichero comenzará inmediatamente después del área reservada, y usará cuantos sectores en secuencia necesite. El siguiente seguirá inmediatamente, y así sucesivamente; sin embargo, una vez que el disco se ha usado y numerosos ficheros han sido añadidos y borrados, el espacio libre en el disco no tendrá trozos tan grandes como para escribir los ficheros completos que siguen, por lo que éstos se fragmentarán. Por tanto será preciso una lista para mostrar qué sectores están libres, y esto lo proporciona la tabla de localización. Esta consiste en una tabla con una entrada correspondiente a cada sector del disco. Cuando el sector está libre, la correspondiente entrada en la tabla de ubicación de ficheros (FAT) se pone a cero. Cuando se escribe un fichero en disco, la entrada correspondiente al sector, usa un conjunto de otros valores que veremos a continuación.

El S.O. ahora ya puede encontrar el número de sectores o 'clusters' libres que sean necesarios, pero no puede obtenerlos todos consecutivos. Necesitamos un método de listado de todos los sectores usados para un fichero y el orden en que se encadenan. De nuevo, la solución es la FAT; cada vez que se escribe el sector, seleccionamos su entrada en la FAT para dar la dirección del sector donde vamos a escribir la siguiente parte del fichero. De este modo, todos los sectores que usamos están encadenados por la FAT, marcamos el último sector del fichero con una entrada especial, en lugar de la dirección del sector en la FAT. Los sectores empleados se mantienen como una lista enlazada.

Esta es una descripción simplificada de como un S.O. típico usa el formato lógico del disco. En la práctica existen varios procedimientos ya que las entradas de la FAT pueden referirse a 'clusters', o sea, a varios sectores en lugar de a un único sector; por otra parte el disco puede dividirse en varias particiones, lo que se muestra al usuario como discos separados, y finalmente, el directorio puede almacenar información adicional sobre cada fichero y puede haber una jerarquía de directorios y subdirectorios.

2.19 ORGANIZACIÓN DEL DISCO EN EL S.O. DOS

Formato de los discos (512 bytes por sector):

- 360 Kb (5 y 1/4 pulgadas): Baja densidad: 40 pistas y 9 sectores por pista
- 720 Kb (3 y 1/2 pulgadas): Baja densidad: 80 pistas y 9 sectores por pista
- 1.2 Mb (5 y 1/4 pulgadas): Alta densidad: 80 pistas y 15 sectores por pista
- 1.44 Mb (3 y 1/2 pulgadas): Alta densidad: 80 pistas y 18 sectores por pista
- Discos duros: normalmente 17 sectores por pista y número de pistas según capacidad

En los disquetes, el primer sector (pista 0, sector 1) contiene el nombre de registro de arranque que es un pequeño programa que permite al ordenador manejar unidades de disco, al menos lo suficiente como para leer otras partes del DOS. Seguidamente aparecen dos copias de la tabla de ubicación de ficheros, que es una especie de índice sobre la distribución de espacios dentro del disco (la segunda copia existe por razones de seguridad). A continuación, se incluye una copia del directorio raíz, que es una lista de los ficheros y referencias a subdirectorios, con indicación del lugar del disco donde comienzan. Por último, aparecen dos pequeños programas DOS, que se leen al comienzo, y que otorgan al ordenador la capacidad necesaria para buscar y cargar el COMMAND.COM, que es el intérprete de comandos del sistema operativo en disco (DOS).

Los discos fijos poseen un registro de arranque principal que contiene una tabla de partición, que permite dividir el disco entre varios sistemas operativos. La tabla de partición contiene información sobre la partición DOS al comienzo del disco, y el primer registro de dicha partición que contiene el registro de arranque DOS. Por lo demás, la partición se organiza igual que en los disquetes.

2.19.1 Estructura lógica del disco

Sea cual sea el disco que se utilice, los discos del DOS están todos formateados lógicamente de la misma forma: las caras, las pistas y los sectores están identificados utilizando la misma notación, y ciertos sectores están siempre reservados a programas e índices especiales que utilizan el DOS para gestionar las operaciones del disco. Las pistas están numeradas del 0 (la exterior) hasta n (la interior).

El BIOS ('Basic Input Output System') localiza los sectores en un disco mediante un sistema de coordenadas en tres dimensiones, compuesto por un número de pista, un número de cara

(número de cabeza) y un número de sector. La secuencia comienza con el primer sector del disco: sector 1 pista 0 cara 0.

Se puede apuntar a un sector determinado, bien por sus coordenadas en tres dimensiones, bien por su orden secuencial. Todas las operaciones de la ROM-BIOS utilizan las coordenadas en tres dimensiones para localizar un sector. Todas las operaciones del DOS y herramientas tales como el debug, utiliza la notación secuencial del DOS.

2.19.2 Organización de los discos

Además de dividir el disco en sectores, el DOS realiza otras operaciones cuando formatea un disco. A continuación veremos la distribución del espacio del disquete:

El proceso de formateo divide los sectores de un disco en cuatro secciones para cuatro usos diferentes. Las secciones, en el orden en que están almacenadas, son: el registro de puesta en marcha, la tabla de localización de ficheros (FAT), el directorio y el espacio de datos. A continuación se hace una breve descripción de cada una de ellas.

- El registro de puesta en marcha:

Es siempre un sector único situado en el sector 1, pista 0, cara 0. El registro de puesta en marcha contiene, entre otras cosas, un cierto programa para comenzar el proceso de carga del sistema operativo. Todos los disquetes contienen el registro de puesta en marcha, aunque no tengan el sistema operativo. Aparte del programa de puesta en marcha o autoarranque, el contenido exacto del registro varía de un formato a otro.

- Tabla de localización de ficheros (FAT):

Está situada a continuación del registro de puesta en marcha, comenzando normalmente en el sector 2, pista 0, cara 0. La FAT contiene el registro oficial del formato del disco y los mapas de localización de los sectores utilizados por los ficheros. El DOS utiliza la FAT para guardar un registro de la utilización del espacio de datos. Cada entrada de la tabla contiene un código específico para indicar el espacio que está siendo utilizado, el que está disponible y el espacio que está defectuoso. Al utilizarse la FAT para controlar todo el área utilizable de almacenamiento de datos, se conservan dos copias idénticas de ella, en previsión de que alguna se dañe. Ambas copias de la FAT pueden ocupar tantos sectores como necesiten: 2 ó 4 en discos flexibles y más de 80 en discos duros. En todos los discos duros, el tamaño de la FAT varía con el tamaño de la partición.

- El directorio de ficheros:

Es el siguiente elemento del disco. Se utiliza como tabla de contenidos, identificando cada fichero del disco como un elemento de directorio que contiene cierta cantidad de información, como el nombre y tamaño de los ficheros. Una parte de la entrada es un número que apunta al primer grupo de sectores utilizados por el fichero (este número es también la primera entrada de este fichero en la FAT). El tamaño del directorio varía según el formato del disco.

- El espacio de datos:

Ocupa la mayor parte del disquete (desde el directorio al último sector), se utiliza para almacenar datos realmente, mientras que las otras tres secciones se utilizan para organizar el espacio de datos. Los sectores del espacio de datos están organizados en unidades conocidas como 'clusters'. El tamaño de un "cluster" varía según el formato. Pueden aparecer 'clusters' que contengan varios sectores.

2.19.3 El registro de arranque (BOOT)

El programa de autoarranque consiste principalmente en un corto programa, en lenguaje máquina, que activa el proceso de carga de DOS en memoria. Para realizar esta tarea, el programa comprueba primero si el disco está formateado por el sistema (si contiene los ficheros IBMBIO.COM y IBMDOS.COM en las versiones de IBM ó MSBIO.COM y MSDOS.COM en la

versión de Microsoft) y entonces procede en secuencia. Normalmente, en la mayoría de los formatos de disco se encontrarán en el registro de arranque algunos parámetros claves que comienzan en el cuarto byte. Estos parámetros son parte del bloque de parámetros del BIOS utilizados por el DOS para controlar cualquier dispositivo tipo disco. El resto del programa de arranque empieza en los primeros tres bytes (0, 1 y 2) y continua en los bytes siguientes al bloque de parámetros de BIOS (Tabla 2.1).

Offset	Longitud	Descripción
3	8 bytes	ID del sistema (ej. IBM 3.1)
11	1 palabra	Nº de bytes por sector (ej. 512=0200 hex)
13	1 byte	Nº de sectores por 'cluster' (ej. 01 ó 02)
14	1 palabra	Nº de sectores reservados al principio: 1 para disquete
16	1 byte	Nº de copias de la FAT: 2 para disquete
17	1 palabra	Nº de elementos del directorio raíz (ej. 64 ó 112)
19	1 palabra	Nº total de sectores del disco (ej. 720 para el D-9)
21	1 byte	de formato (ej. FF, FE, FD o FC)
22	1 palabra	Nº de sectores por FAT (ej. 1 ó 2)
24	1 palabra	Nº de sectores por pista (ej. 8 ó 9)
26	1 palabra	Nº de caras (cabezas) (ej. 1 ó 2)
28	1 palabra	Nº de sectores especiales reservados

Tabla 2.1 Parámetros del registro de arranque

2.19.4 Tabla de localización de ficheros

Hay que distinguir entre como está organizada la FAT, que es relativamente simple e inmediato, y como está almacenada en el disco, lo cual es más complejo. Cada copia de la FAT ocupa dos sectores en los formateos de 9 sectores por pista y siete sectores en los formateos de 15 (Tabla 2.2).

Hay dos formatos para la FAT: uno de 12 bits y otro de 16 bits. El formato de 12 bits es el más extendido y el más complicado. La FAT está organizada como una tabla de hasta 4096 números, con un elemento para cada 'cluster' en el espacio de datos. El número que contiene cada elemento indica el estado y uso del 'cluster' correspondiente. Si el elemento de la FAT es 0, se indica que el 'cluster' está libre y disponible para su uso. Si el elemento de la FAT contiene 4087 (FF7 hex) el 'cluster' está declarado como inutilizable por un error de formateo. Los valores de la FAT del 4081 al 4086 (FF1 al FF6 hex) se reservan también para señalar la imposibilidad de utilizar un determinado 'cluster', pero no se utilizan.

Elemento de la FAT	Valor		Significado
	Dec.	Hex.	
0	253	FD	El disco es doble cara, doble densidad
1	4094	EFE	Entrada no utilizada, disponible
2	3	003	El siguiente 'cluster' del fichero es el 'cluster' 3
3	5	005	El siguiente 'cluster' del fichero es el 'cluster' 5
4	4087	FF7	El 'cluster' es no utilizable: pista mala
5	6	006	El siguiente 'cluster' del fichero es el 'cluster' 6
6	4095	FFF	Último 'cluster' del fichero y final de esta cadena de atribución de espacio
7	0	0	Entrada no utilizada

Tabla 2.2 Cadena de atribución de espacio de un fichero en la tabla de atribución de ficheros

Los 'clusters' están numerados por orden desde el 2 hasta un número que sea superior en una unidad al número de 'clusters' del disco. Una entrada en la FAT de 12 bits que contenga cualquier número entre 2 y 4010 (02 y FF0 en hex) indica que el 'cluster' correspondiente está siendo utilizado por un fichero. Un valor de la FAT de 4095 (FFF hex) indica que el correspondiente 'cluster' contiene la última parte de los datos de un fichero. Unos valores entre 4008 y 4094 (FF1 al FFE) tendrán el mismo significado, pero no se utilizan. El elemento del directorio del fichero contiene el número del 'cluster' de comienzo y las entradas de la FAT designan los demás 'clusters' utilizados y el final del fichero. Cuando un fichero es borrado, todos los elementos de la FAT que determinan su cadena de localización de espacio son marcados como disponibles (puestos a cero); pero los datos del fichero en el espacio de datos no sufren modificación alguna, y la mayor parte de la información del elemento se conserva. Aunque el valor de la FAT sea simple, la grabación se hace de una forma más compleja. El rango de números de 'cluster' está definido de forma que los elementos de la FAT sean 4095 (FFF hex) o menos. Esto hace posible cada elemento de tres dígitos hexadecimales en 12 bits o un byte y medio. Los elementos de la FAT se agrupan por pares, ocupando cada par tres bytes. Los tres bytes se codifican de la siguiente forma: si un par de elementos de la FAT está formado por 123 y 456 hex, los tres bytes que los contienen serían en hexadecimal 23-61-45. En sentido inverso, si los tres bytes son AB-CD-EF, los dos valores de la FAT son DAB y EFC. Dado cualquier número de "cluster" se puede encontrar el valor de la FAT multiplicando el número de 'cluster' por tres, dividiendo por dos y utilizando el número completo del resultado como un desplazamiento de la FAT. Cogiendo una palabra de esa ubicación, se tendrán los tres dígitos hexadecimales de elemento de la FAT, más un dígito hexadecimal extraño que se puede ignorar. Si el número de 'cluster' es impar, se desecha el dígito de mayor orden, si es par el dígito de menor orden. El valor obtenido de esta manera es el número del siguiente 'cluster' del fichero, a menos que sea FFF, que indica el último 'cluster' de un fichero. Los detalles reseñados hasta ahora son útiles para la FAT de 12, que pueden alojar hasta 4010 'clusters'. Si un formato de disco tiene un número superior de 'clusters', es necesario utilizar una FAT de 16 bits.

2.19.5 El directorio

Los directorios de los discos se utilizan para almacenar la mayor parte de la información básica sobre los ficheros contenidos en el disco, incluyendo el nombre de los ficheros, su tamaño, el comienzo de elemento de la FAT, la hora y fecha en que fueron creados y unas pocas características especiales del disco. La única información que no contiene el directorio es la localización exacta de los 'clusters' individuales que componen el fichero (éstos están almacenados en la tabla de localización de ficheros).

Hay un elemento en el directorio para cada fichero del disco, incluyendo los ficheros de subdirectorio y la etiqueta de identificación de volumen del disco. Cada uno de los elementos tiene 32 bytes, por lo que un sector del directorio puede almacenar 16 elementos. Cada entrada de 32 bytes del directorio está dividida en ocho campos (tabla 2.3):

- Campo 1: el nombre del fichero

Los ocho primeros bytes de cada elemento del directorio contienen el nombre del fichero, almacenado en formato ASCII. Si el nombre del fichero tiene menos de ocho caracteres, se completa por la derecha con espacio en blanco. Las letras deberán ser mayúsculas ya que las minúsculas no son reconocidas correctamente. El primer byte de los directorios que no se usan es 00. Cuando se borra un fichero, solo quedan afectadas dos cosas del disco. Al primer byte se le asigna el valor E5 hex y la cadena de utilización de espacio en fichero se borra en la FAT. El resto de la información del directorio acerca del fichero se mantiene. La información perdida puede ser recuperada mediante métodos adecuados, siempre que el elemento del directorio no se haya utilizado por otro fichero. Se advierte que cuando es necesario crear un nuevo elemento de directorio, el DOS utiliza el primero disponible, reciclando rápidamente la entrada correspondiente a un fichero ya borrado, haciendo imposible la recuperación. El tercer código que se puede encontrar en el byte del nombre de fichero es el carácter punto, 2E hex, que se utiliza para

especificar un subdirectorio. Si el segundo byte es también 2E hex, indica que se trata del directorio padre del directorio en uso, en cuyo caso el campo 'cluster' de comienzo contiene el número de 'cluster' del directorio padre.

Campo	Offset	Descripción	Tamaño (bytes)	Formato
1	0	Nombre del fichero	8	Caracteres ASCII
2	8	Extensión del nombre de fichero	3	Caracteres ASCII
3	11	Atributo	1	Bit codificador
4	12	Reservado	10	No utilizados; ceros
5	22	Hora	2	Palabra, Codificada
6	24	Fecha	2	Palabra, Codificada
7	26	Comienzo de entrada a la FAT	2	Palabra
8	28	Tamaño del fichero	4	Entero

Tabla 2.3 Partes de un elemento directorio

- Campo 2: extensión del nombre del fichero.

Después del nombre de fichero se encuentra la extensión estándar del nombre de fichero, en formato ASCII. Son tres bytes y como el nombre del fichero, se completa con espacios en blanco si consta de menos de tres caracteres, hasta alcanzar su número.

- Campo 3: atributos del fichero.

El tercer campo de la entrada de directorio consta de un byte, y cada uno de sus bits se utiliza para caracterizar el elemento de directorio.

bit 0: caracteriza a un fichero como de solo lectura

bits 1 y 2: caracteriza a los ficheros como ocultos o del sistema.

bit 3: indica que el elemento de directorio es una etiqueta de identificación (ID de volumen del disco). Las etiquetas solo son reconocidas correctamente en directorio raíz, y solo utiliza unos pocos de los ocho campos disponibles en el elemento.

bit 4: el atributo de subdirectorio, sirve para identificar elementos de directorio que identifican los subdirectorios. Dado que los subdirectorios están almacenados en el disco como ficheros de datos ordinarios, necesitan un elemento de directorio propio. Estos elementos utilizan todos los campos del directorio, excepto el campo de tamaño del fichero, que toma el valor cero. El tamaño real de un subdirectorio se determina siguiendo su cadena de localización, que hay que buscar en la FAT.

bit 5: el atributo de archivo, fue creado para facilitar la realización de copias de seguridad de algunos ficheros que pueden estar almacenados en un disco duro. Este bit está a cero en todos los ficheros que no han cambiado desde la última copia de seguridad. Y el bit está normalmente a uno en todos los ficheros de un disquete. El atributo de archivo no resulta particularmente útil en los disquetes.

- Campo 4: reservado.

Diez bytes reservados para posibles usos futuros

- Campo 5: la hora

Contiene un valor de dos bytes que señala la hora en la que fue creado o sufrió su último cambio el fichero.

- Campo 6: la fecha.

Contiene un valor de dos bytes que indica la fecha en que se creó el fichero, o en que se modificó por última vez.

- Campo 7: número de 'cluster' de comienzo.

Este campo consta de dos bytes que indican el número de 'cluster' de comienzo del espacio de datos del fichero. Actúa como el punto de entrada en la cadena de localización del fichero en la FAT.

- Campo 8: tamaño del fichero.

Indica el tamaño del fichero en bytes. Está codificado como un entero sin signo de cuatro bytes. Normalmente este número indica el tamaño exacto del fichero. Pero en algunos ficheros, sobre todo los generados por procesadores de texto, los cuales trabajan con bloques de bits (128 generalmente), puede haber alguna diferencia. En cualquier caso, cuando el DOS está leyendo un fichero, establece el final del fichero cuando lee todo el fichero, según su tamaño, o cuando llega al final de la cadena de localización de la FAT (denotado por FFF hex), sea cual sea el que aparezca primero.

2.19.6 El espacio de datos

El espacio se otorga a los ficheros a medida que lo necesitan, un 'cluster' cada vez. Las últimas versiones del DOS añaden siempre nuevos 'cluster' mediante reglas complicadas que no veremos. En la mayoría de las ocasiones, cada fichero se almacena en un bloque de espacio continuo. Sin embargo, un fichero puede dividirse en varios bloques no contiguos, especialmente cuando se añade información a uno ya existente, o cuando se almacena un nuevo fichero en el espacio dejado por un fichero borrado.

2.20 EL ALMACENAMIENTO ÓPTICO

Podemos definir el almacenamiento de datos diciendo que ello significa alterar alguna propiedad del conjunto de elementos de almacenamiento de tal forma que dicha alteración representa el dato a ser almacenado; posteriormente en base a esa propiedad, reconstruiremos el dato original a partir de ella. El almacenamiento óptico es una clase de almacenamiento de datos en la que se detecta esta propiedad por medios ópticos. 'Alterar' y 'detectar' corresponden en términos de computación, a escribir y leer. Deliberadamente no hemos utilizado el escribir datos por medio ópticos. De hecho, un dispositivo de almacenamiento óptico actual no escribe datos por procesos puramente ópticos, aunque tal proceso exista. El proceso usado en la práctica utiliza un rayo de luz para escribir, pero el cambio de las propiedades ópticas del medio es causado por efectos de calentamiento de ese rayo. Esto, por supuesto, no es aplicable a medios de sólo lectura, que son copiados desde un disco maestro por medios puramente mecánicos (mediante prensado).

Hay varias propiedades ópticas que pueden utilizarse para almacenamiento de datos, de las cuales la más simple es la reflectividad de la superficie del medio. Es fácil detectar los cambios en la reflectividad por la brillantez de la luz en la superficie del medio detectando la luz reflejada con un fotodetector. Sin embargo, esto es un proceso activo, al contrario que en los medios magnéticos donde simplemente se detecta el voltaje inducido en la cabeza. En el caso que ahora nos ocupa, estamos utilizando rayos de luz tanto para escribir como para leer en el medio óptico, y queremos que la operación de lectura no sea destructiva. Para conseguir esto, leemos con un rayo menos potente que el usado para escribir, y el medio debe tener bien definido el umbral para que un rayo de menor potencia en la lectura no cambie la información almacenada. Sin embargo, al ser reescrito muchas veces, el material puede perder sus propiedades iniciales y por lo tanto, no es

sorprendente que el mayor inconveniente en la tarea de introducir los almacenamientos ópticos en el mercado haya sido el desarrollo de medios satisfactorios.

Para medios reescribibles, aunque no para WORM (Write Once, Read Many times), hay otra dificultad, y es la de poder asegurar que los efectos que produce el calentamiento del material, cuando escribimos, son completamente reversibles. El mayor problema en el desarrollo de medios ópticos reescribibles es la fatiga del medio que limita el número de modificaciones que admite. Muchos materiales que en principio parecían útiles tuvieron que ser descartados porque tenían un límite bajo en el número de veces que los datos podían ser borrados y reescritos. El límite, para alguno de los materiales, es ahora mucho más alto y puede utilizarse de forma aceptable para la mayoría de los propósitos. Este problema de la fatiga retrasó la llegada de los almacenamientos ópticos reescribibles al mercado hasta cuatro o cinco años después de que los dispositivos WORM estuvieran disponibles en laboratorio.

La mayoría de los medios WORM y reescribibles dependen de escrituras termo-ópticas y de cambios en la reflectividad, pero hay una importante excepción. Esta es la grabación magneto-óptica, usualmente abreviada como MO y que es el proceso usado en el primer dispositivo óptico reescribible que llegó al mercado. La escritura en medios MO no sólo requiere un efecto de calentamiento del rayo de luz, sino que además, como su propio nombre indica, se aplica un campo magnético al punto que es calentado. Mediante un laser se calienta una zona muy pequeña de la superficie por encima de una temperatura conocida como punto de Curie. Por encima de esta temperatura los dominios magnéticos del medio pueden ser orientados bajo la acción de un campo magnético externo. Cuando el material se enfría por debajo de esta temperatura de Curie, el material retiene la orientación de los dominios magnéticos y por lo tanto la magnetización y los datos inicialmente grabados, aunque se someta nuevamente al campo magnético. La lectura depende de la dirección de polarización del haz reflejado en lugar de la intensidad del rayo de luz reflejado. Sin embargo, este cambio de polarización es convertido por una óptica en un cambio de intensidad de la luz que llega al fotodetector, por lo que, eléctricamente, el método de lectura no es muy diferente. El fenómeno físico por el cual una superficie magnetizada refleja la luz con un eje distinto de polarización según la orientación del campo magnético se conoce como efecto Kerr magneto-óptico.

2.20.1 El sistema óptico

Los discos ópticos son similares a los discos magnéticos en que las pistas son leídas y escritas por una cabeza que se mueve aproximadamente de manera radial para dar acceso a varias pistas. Sin embargo, el diseño de la cabeza es totalmente diferente al de la cabeza magnética. La cabeza óptica es mucho más abultada y más cara; por lo tanto, los cilindros propuestos en los discos magnéticos y que suponen la existencia de múltiples superficies y por lo tanto múltiples cabezas de lectura son menos apropiados para los discos ópticos, que tienen una sola cabeza y por lo tanto acceden a una sola superficie de un disco a la vez. Estos pueden tener una única superficie activa, pero en algunos casos, como sucede con el DVD (Digital Versatile Disk), el disco es de doble cara. En este caso, debe ser retirado del dispositivo y se le da la vuelta manualmente para tener acceso a la segunda cara. Lectores con más de una cabeza aparecerán posiblemente en el futuro, pero por ahora son minoritarios.

El CD-ROM cuenta con una única pista en espiral que mide unos 5 km dividida en 270.000 sectores en los discos de 60 minutos, en 330.000 en los de 74' y en 360.000 en los de 80'. Los sectores son de 2352 bytes de los cuales los 12 primeros son de sincronización, los tres siguientes de cabecera, a continuación vienen los 2048 bytes de datos y el resto son para los códigos de detección y corrección de errores. La separación entre dos vueltas adyacentes, es decir el paso de vuelta, es de 1,6 μm . Aunque este tipo de dispositivos tienen una única pista espiral, normalmente se denomina pista a cada una de las vueltas que incorpora, por analogía con las pistas en un disco magnético. Esta espiral continua está formada por abultamientos de la superficie denominados

"pits" que sobresalen por encima de la superficie que recibe el nombre de "land". El ancho de estos abultamientos es de $0.5\text{ }\mu\text{m}$. La altura de estos abultamientos es de $1/4$ de la longitud de onda del láser empleado para la lectura. De esta forma, se producirá interferencia destructiva, ya que la diferencia entre los rayos de luz que se reflejan en un "pit" y en un "land" se diferenciarán justamente en media longitud de onda. La separación mínima entre un flanco y otro, es decir la longitud de un "pit" (o "land") o la separación entre "pits" (o "lands") es de $0.833\text{ }\mu\text{m}$ y la separación máxima de $3.054\text{ }\mu\text{m}$. Esto corresponde a las distancias mínima y máxima sin transición. Estas separaciones máximas y mínimas vienen impuestas por el código, que es un tipo de código RLL-3,11 que se conoce como modulación de ocho a catorce (EFM= Eighth to Fourteen Modulation), ya que a grupos de ocho bits de datos se les asignan grupos de catorce bits, de forma análoga a como se hacía con el código RLL-2,7. Por último, el diametro del punto luminoso sobre la superficie es de unos $1.5\text{-}1.6\text{ }\mu\text{m}$ de diámetro. La figura (2.22 a) muestra estas dimensiones.

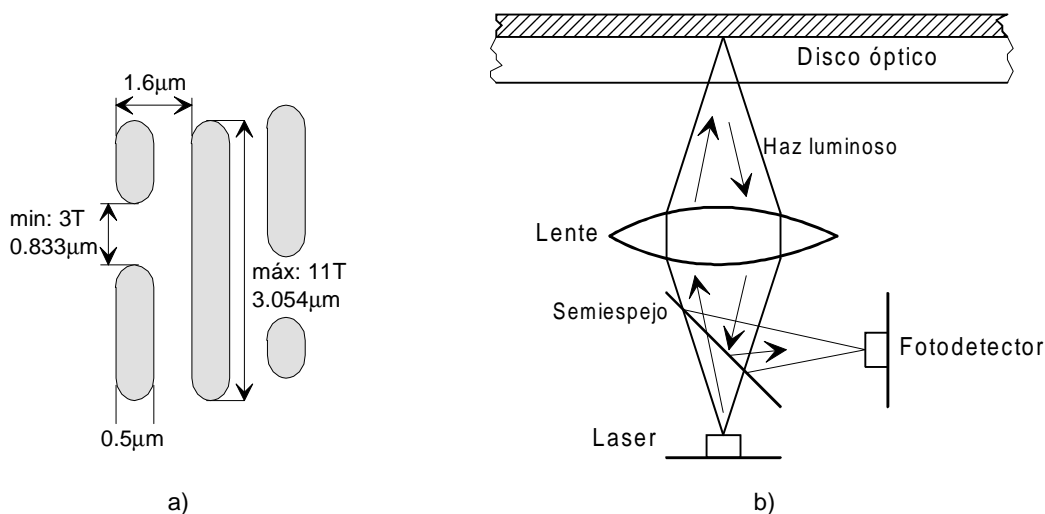


Fig. 2.22 a) Dimensiones típicas de un disco óptico
b) Sistema óptico básico de un lector de disco óptico

El sistema óptico está compuesto por varios elementos, y en muchos dispositivos todos ellos son movidos de pista a pista. En otros casos, la mayoría de estos componentes están fijos y sólo partes del sistema se mueven. Aunque la terminología varía, será conveniente ver la cabeza únicamente como aquellos componentes que se mueven.

Los componentes básicos del sistema óptico se muestran en la figura (2.22 b). Dicho sistema consiste en un semiconductor láser para generar un rayo de luz, un fotodetector para detectar la luz reflejada de la superficie del disco, una lente objetivo para enfocar el rayo láser a un punto en la superficie activa del medio, y el semiespejo separador del haz que dirige la luz reflejada desde ese punto (a través de la lente objetivo) al fotodetector. En los sistemas prácticos hay componentes adicionales, aunque estos varían de un sistema a otro. Vamos a ver a continuación cada uno de los componentes básicos.

El almacenamiento óptico es descrito a veces como "almacenamiento láser"; la palabra "láser" se ve como una potente herramienta de venta. De hecho, el láser no es la única fuente de luz posible. Algunos de los primeros experimentos en almacenamiento óptico usaron otras fuentes de luz, y otras nuevas pueden usarse en el futuro. Por otra parte, unos pocos dispositivos experimentales dependerán de efectos ópticos que pueden ser producidos sólo por láser. En la práctica, todos los dispositivos de almacenamiento óptico que están en el mercado usan los láser como fuente de luz. El láser semiconductor, aunque en principio no es esencial, ha proporcionado una fuente de luz muy adecuada. Se ha vuelto muy fiable y razonablemente barato. Estas características se deben fundamentalmente a su amplio uso por la industria de los sistemas de discos compactos de audio. En realidad, muchos de los componentes usados en el almacenamiento

óptico son derivados de esta industria. Otro tipo de láser es el láser de gas, que fue usado en el video disco y en algunos de los primeros almacenamientos ópticos desarrollados. El láser de gas es más potente que el de semiconductor, pero es más grande, menos fiable, más caro y más difícil de modular por la electrónica de control.

El láser semiconductor tiene tres propiedades que lo hacen una fuente de luz adecuada para el almacenamiento óptico. En primer lugar, emite luz de una sola longitud de onda lo que simplifica el diseño de los otros componentes ópticos y del medio. Segundo, la luz emitida por él se concentra en un estrecho rayo (aunque no tan estrecho como el del láser de gas), lo que simplifica de nuevo el sistema óptico. Por último, la potencia del rayo puede modularse fácilmente por una señal eléctrica. Hay unas propiedades que son interesantes en otras aplicaciones, pero no es necesario discutir las aquí. Los láseres utilizados en almacenamiento óptico tienen una potencia de salida bastante baja, de 10 a 30 miliwatios. La luz emitida está en el infrarrojo cercano, porque los láseres con estas longitudes de onda son fáciles y económicos de fabricar. Los láseres azules podrían ser mejores: la densidad en la que los datos pueden ser almacenados depende, entre otras cosas, de la longitud de onda del láser, y los láseres que emiten luz azul tienen una longitud de onda de alrededor de la mitad del láser infrarrojo. Los láseres semiconductores azules han sido fabricados, pero aún no han sido desarrollados hasta el punto de poder ser usados en dispositivos prácticos. Al tener una longitud de onda casi la mitad que el láser rojo empleado en el DVD, un dispositivo que emplease un láser azul podría multiplicar por cuatro la capacidad de estos dispositivos, al igual que los DVD aumentan considerablemente su capacidad frente a los CD-ROM gracias a que emplean luz roja, en lugar de infrarroja, que tiene una longitud de onda menor y por lo tanto permite una mayor densidad de almacenamiento.

La necesidad del fotodetector es simplemente producir una señal eléctrica que se corresponda con la cantidad de luz que incide en él. Tales dispositivos están disponibles desde hace muchos años. En la práctica, este dispositivo es más complejo de lo que esta descripción implica.

Al igual que sucede con los láseres, no hay un único tipo de lentes objetivo. Estas tienen que trabajar sólo con luz monocromática, por lo que su diseño es muy fácil. Tienen una gran apertura (en otras palabras, su diámetro es grande con respecto a su distancia focal) para hacer el mejor uso de la cantidad de luz disponible. Es posible fabricar estas lentes con una distancia focal suficientemente grande como para que el punto más cercano de la lente a la superficie del disco sea amplio. Esta separación es típicamente alrededor de un par de milímetros, que es una distancia muy grande comparada con la altura de vuelo de la cabeza en los discos magnéticos. Esto tiene dos beneficios importantes. El primero es que es fácil de diseñar las unidades de disco para que el disco pueda ser quitado y reemplazado por otro, lo cual es imposible con los discos magnéticos de alta densidad. Segundo, tenemos una región para poner una fina capa sobre la cara activa del disco para protegerlo, por lo que cualquier partícula de polvo o arañazo menor en la superficie externa del disco, al estar fuera del foco, no tiene una influencia significativa en la lectura y escritura. Esto hace reducir el contraste de la imagen reflejada en el fotodetector y por tanto la relación señal-ruido del sistema. Por este motivo se necesita bastante polvo o arañazo para tener un efecto significativo en la integridad del dato. Esto hace al disco óptico un medio muy robusto en contraste con los discos magnéticos y por extensión las cintas magnéticas.

El último componente vital del sistema óptico es el separador del haz. Este dispositivo era bien conocido antes de que fuera aplicado al almacenamiento óptico. Su propósito es dividir la luz reflejada en su camino saliente para que llegue al fotodetector del láser. Hay varios dispositivos que pueden hacer esto. El más simple es un semiespejo, el cual tiene un revestimiento reflectante que lo hace parcialmente transparente. Este no es muy efectivo puesto que solo una parte del rayo retornado es desviada y algunos de los rayos que salen de la lente son dispersados y se pierden. Los dispositivos prácticos están normalmente basados en prismas, diseñados de tal forma que el rayo retornante sea reflejado hacia el detector con lo que se aprovecha toda la intensidad luminosa.

Por último, señalemos que pueden añadirse más componentes para soportar sistemas servo que mantiene el punto de foco en la pista de datos para mejorar el tiempo de acceso del dispositivo, o simplemente para permitir un formato más conveniente.

2.20.2 Seguimiento de la pista

La cabeza óptica, como la cabeza magnética, debe seguir la pista de datos escrita en el disco. Vimos que en algunos discos magnéticos este seguimiento puede hacerse por cómputo o cálculo estimado sin realimentación. Esto no es posible en unidades de discos ópticos porque las pistas están mucho más cercanas. Además, los discos son normalmente intercambiables; la localización precisa de la pista puede variar de un dispositivo a otro, y los discos pueden tener las pistas ligeramente excéntricas. Una complicación añadida es que la mayoría de los discos ópticos no usan pistas concéntricas separadas, sino que los datos están ordenados a lo largo de una espiral continua, aunque es normal considerar cada una de las vueltas de la espiral como una pista. Por todo esto las servopistas son siempre necesarias, y usan las servotécnicas embebidas (indicador de la pista en la misma pista) más que una superficie y cabeza separada, ya que además las cabezas ópticas son considerablemente más caras que las magnéticas y por lo tanto no resulta rentable.

La separación entre pistas, como ya se adelantó al comienzo de este apartado, es tan sólo de $1.6\text{ }\mu\text{m}$, que es mucho menor que la precisión con la que pueden fabricarse de forma rentable el plato del reproductor o el orificio central del disco. En un reproductor típico, si mantenemos fija la cabeza, debido a las excentricidades del sistema o del propio disco, durante una vuelta pasarán por delante de ella varias pistas. Es decir la excentricidad o irregularidad en el paso de pista es muy superior a la separación entre ambas por lo que se hace necesario un sistema de seguimiento con una dinámica mucho más rápida que la velocidad de giro del disco. Hay varios métodos para conseguir esto.

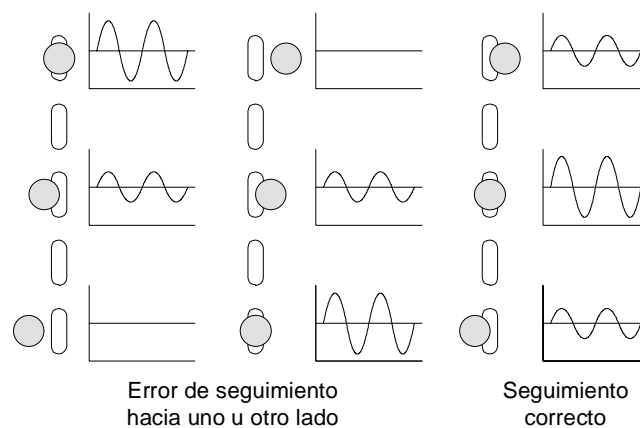


Fig. 2.23 El método de tres puntos para el seguimiento de pista

En el método de tres puntos, se enfocan tres haces de luz sobre la pista del disco, uno de los cuales es desplazado ligeramente a un lado de la pista y el otro al otro lado de la línea central de la misma, tal y como se muestra en la figura (2.23). Cuando el haz central, que es el que realmente recoge la información, está en el centro, los otros dos haces estarán parcialmente sobre la pista lo que producirá en sus respectivos fotodetectores (hay un fotodetector para cada haz) una pequeña señal oscilante. Si por el contrario la cabeza se aleja del centro de la pista, antes de que el haz principal la abandone lo habrá hecho alguno de los otros, según el lado hacia el que se produzca la salida. En ese momento, el fotodetector correspondiente no recibirá la señal ondulatoria y se podrá actuar para corregir esa deriva antes de que el haz principal se vea afectado. Los tres haces son generados a partir del haz proveniente del láser con la ayuda de una rejilla de difracción. De hecho,

el haz central es el lóbulo principal de la difracción y los otros dos se corresponden con los lóbulos laterales.

La señal oscilante que reciben los fotodetectores es debida al paso por las marcas propias de la superficie del disco. Estas señales se pasan por sendos filtros paso-bajo para eliminar esta información de canal y obtener el brillo medio. Si los dos haces secundarios producen en sus correspondientes fotodetectores la misma intensidad, el haz principal estará centrado. Si por el contrario hay una diferencia apreciable entre los niveles detectados, el sistema estará a punto de salirse de la pista. El signo de esta diferencia indica en que sentido debe actuarse para recolocar la cabeza en el centro de la pista de datos.

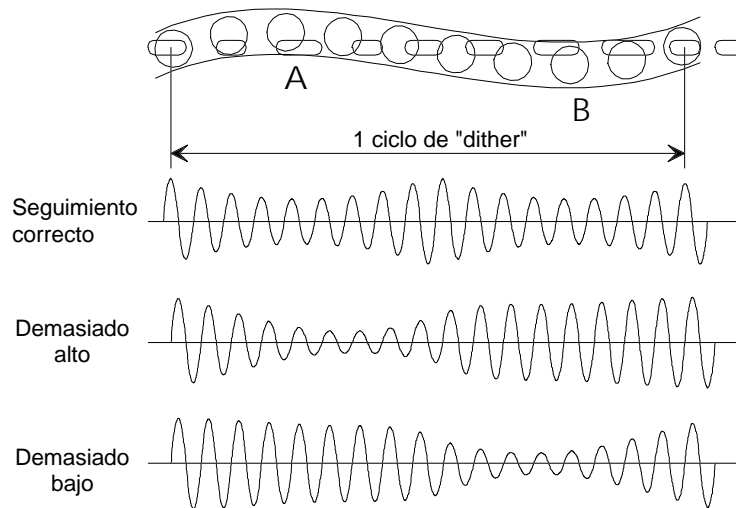


Fig. 2.24 La oscilación aplicada al haz de lectura modula la envolvente de lectura que puede utilizarse para determinar el error de seguimiento

Otro método consiste en superponer una pequeña oscilación al haz de forma que se produzca una modulación de la envolvente de la señal de lectura, que puede detectarse para obtener el sentido del error. Esta oscilación puede generarse mediante la vibración de un espejo en la trayectoria de luz, lo cual permite una alta frecuencia, o bien mediante la oscilación de todo el lector a una frecuencia menor. La figura (2.24) muestra el esquema de funcionamiento de este método. El seguimiento es correcto cuando el nivel de intensidad es similar en los momentos de máximo desplazamiento. Si por el contrario el nivel de señal en los momentos de máximo desplazamiento positivo (A) es muy distinto del nivel en los momentos de máximo nivel de desplazamiento negativo (B), la cabeza debe ser reposicionada.

En los controladores de bajas prestaciones, la cabeza completa puede moverse para seguir la pista por medio de un actuador similar al usado en los discos magnéticos. Pueden utilizarse voice-coil o un motor lineal. El brazo de la cabeza puede moverse en línea, o más comúnmente, en arco. El peso del dispositivo puede reducirse introduciendo un espejo en ángulo recto entre el separador del haz y la lente objetivo, con lo que la parte principal del sistema óptico puede ser puesta en su cara. Esto también hace posible reducir el peso de la cabeza óptica y así mejorar el tiempo de acceso, fijando la parte principal del sistema óptico y poniendo sólo el espejo y la lente objetivo en la cabeza. Es necesario añadir una lente de colimación a la parte fija del sistema de modo que el rayo de luz se mantenga sin dispersarse entre las secciones fijas y en movimiento (Figura 2.25). La desventaja de este diseño es que el alineamiento preciso entre las dos partes del sistema óptico puede ser difícil de mantener.

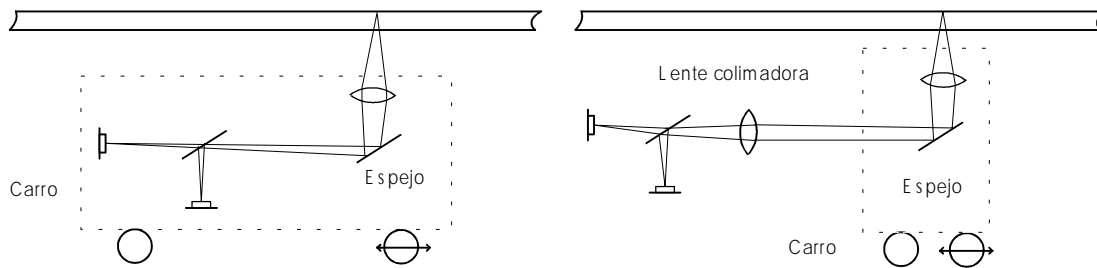


Fig. 2.25 Esquema de montajes alternativos para el sistema óptico.
Izquierda: el láser se mueve con la cabeza. Derecha: el láser está fijo al sistema

Los sistemas de altas prestaciones usan un sistema de seguimiento de pista en dos etapas. El movimiento de la cabeza óptica es como acabamos de ver, pero se le añade un espejo en ángulo recto para habilitar la inclinación y de este modo deflecar el rayo a lo largo del radio del disco. La distancia que puede moverse el rayo de esta forma es muy limitada, porque debe mantener el paso a través de la lente objetivo, aunque en algunos dispositivos esta lente se mueve también o en lugar del espejo. Esto es suficiente para detectar cualquier pérdida de alineamiento o excentricidad, y también movimientos entre pistas adyacentes. Se usan dos mecanismos de servo separados: El primero, usa la información de servo desde el fotodetector, controlando el espejo para mantener el rayo de luz alineado con la pista. El segundo, toma la posición del espejo y mueve la cabeza completa, por lo que el espejo puede retornar a su posición media. La frecuencia de respuesta de este segundo servo es limitada, por lo que no puede seguir la excentricidad de la pista.

2.20.3 Control de enfoque

Al igual que sucede con la excentricidad del sistema plato-disco, el disco no mantiene siempre el mismo plano de giro. Debido nuevamente a imperfecciones del sistema o a deformaciones del disco en forma de alabeos, puede suceder que el haz no enfoque correctamente sobre la superficie de datos. Se hace necesario por tanto un control de enfoque. El posicionamiento de la lente para conseguir el enfoque adecuado, se obtiene con la ayuda de un actuador electrodinámico de bobina móvil, como en el caso de los motores "voice coil" para posicionar las cabezas de los discos duros. En este caso, la lente de enfoque va montada sobre una pequeña bobina suspendida en un campo magnético creado por un pequeño imán. La lente se acercará o alejará en función de la corriente que pase por la bobina (ver figura 2.26).

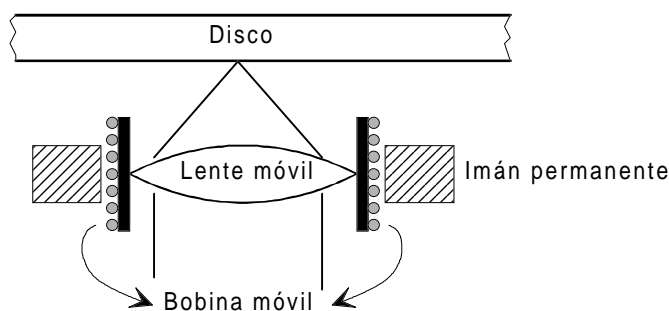


Fig. 2.26 Focalización moviendo la lente objetivo

La figura (2.27) muestra el esquema del control de enfoque conocido como filo de cuchillo. En este caso se requiere un sensor dividido en dos partes. Si el punto focal coincide con el filo del cuchillo, éste tiene poco efecto sobre el haz. Si por el contrario no coincide, apareciendo por delante o por detrás del plano del filo, uno de los dos sensores recibirá mayor cantidad de luz que el otro, porque el filo interrumpe el haz que debe llegar a uno de los dos fotosensores pero no afecta al otro. El error de enfoque se deduce comparando las salidas de las dos mitades del sensor. Esto se puede utilizar para corregir el enfoque. El principal inconveniente de este método es que la

posición lateral del filo es crítica y requiere un ajuste muy preciso. Para solventar esta dificultad se sustituye el filo por un prisma doble (figura 2.28). Las tolerancias en este caso, solo afectan a la sensibilidad, sin producir ningún desplazamiento de enfoque.

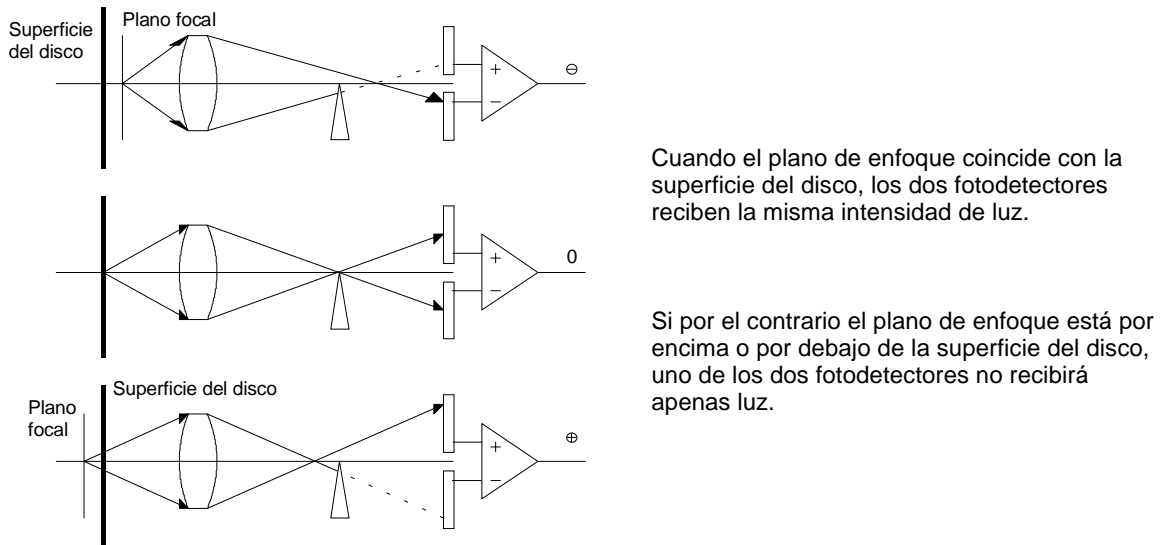


Fig. 2.27 Esquema de funcionamiento del método de control de enfoque denominado de filo de cuchillo. Necesita tan sólo dos sensores pero depende de forma crítica de la correcta posición del filo.

Un método más preciso y que no requiere estos ajustes tan precisos está basado en una lente cilíndrica. Cuando el haz pasa por la lente cilíndrica, como en un eje no tiene curvatura, no afecta al haz en esa dirección, sin embargo en el otro eje tiene el efecto de acortar la longitud focal de todo el sistema. La imagen por tanto será una elipse. Esta elipse será alargada o achatada en este eje según que el punto focal se sitúe por delante o por detrás de esta lente cilíndrica. Si el punto focal coincide con la lente, la alteración será mínima y la imagen será casi completamente circular. En este momento la lente está enfocada. Para determinar la asimetría del punto luminoso se emplea un fotodetector dividido en cuatro cuadrantes (fig 2.29). Si todos los cuadrantes reciben la misma intensidad, la imagen es circular y el sistema está enfocado, si por el contrario dos cuadrantes diagonalmente opuestos reciben más luz que los otros dos el sistema estará enfocando demasiado cerca, y estará enfocando demasiado lejos cuando los que reciben más luz sean el otro par.

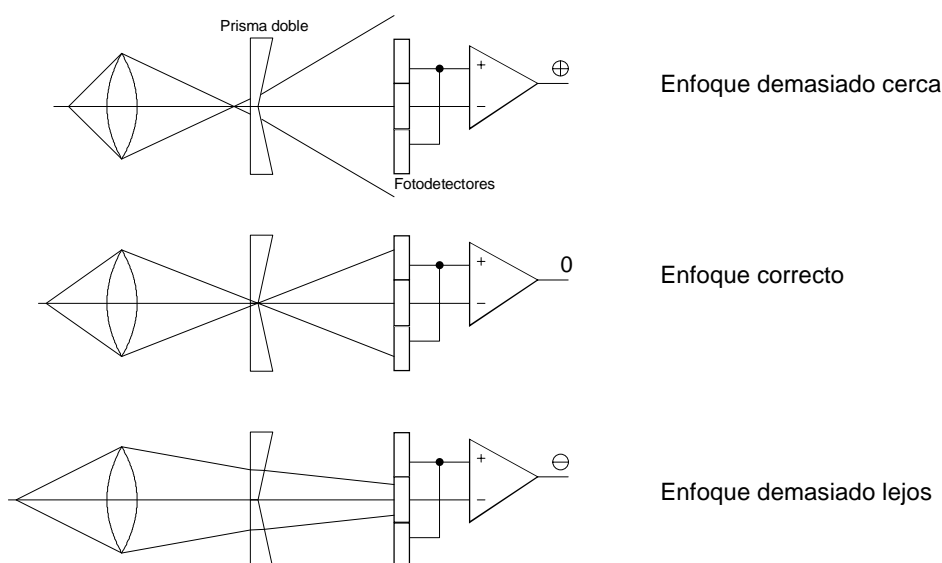


Fig. 2.28 Esquema de un sistema que emplea el método del prisma doble para detectar los errores de enfoque. Este método utiliza tres fotodetectores

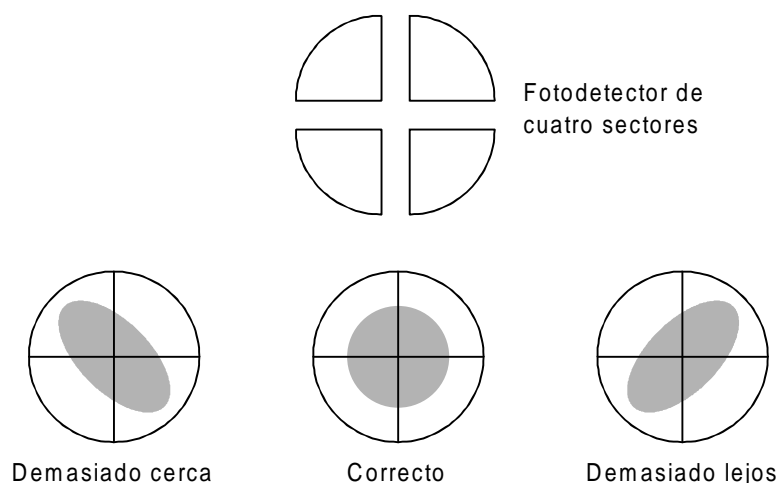


Fig. 2.29 Errores de focalización

2.20.4 Rotación del disco

La mayoría de los discos reescribibles y escribibles una sola vez tienen una velocidad angular constante (Constant Angular Velocity o CAV); este es, por supuesto, el método usado en las controladoras de discos magnéticos, para las que los datos no están densamente empaquetados en las pistas más externas como ya se comentó y donde la velocidad de acceso es un factor crítico. Por otra parte, la potencia requerida para aplicar una cantidad de energía dada a un área del disco, varía con el radio de la pista. El segundo punto no es muy importante en los almacenamientos ópticos, puesto que la potencia del láser puede ser controlada fácilmente.

Sin embargo, algunas controladoras incrementan la capacidad de almacenamiento de datos en cada uno de los discos grabando con velocidad lineal constante (Constant Linear Velocity o CLV). De este modo, la velocidad de rotación del disco se varía con el radio de la pista, por lo que la superficie del disco siempre pasa por la cabeza a la misma velocidad lineal. Así pueden grabarse todas las pistas con la misma densidad (espaciado de bit). Esto incrementa la capacidad del disco en alrededor del 50%.

Hay, sin embargo, un precio que pagar y este es el tiempo de acceso. Antes sólo teníamos que esperar a que la cabeza se moviera a la pista correcta; ahora debemos cambiar la velocidad del disco. La inercia del disco es considerable, y en la práctica, el cambio de velocidad se toma un tiempo considerablemente mayor que el movimiento de la cabeza. En el dispositivo típico CLV, el tiempo de acceso es alrededor de medio segundo. Este es un precio aceptable a pagar para mayores capacidades en algunas aplicaciones, pero no en otras. Muchas controladoras CLV son, por lo tanto, útiles para ambos modos y trabajarán tanto en modo CLV como en CAV, dependiendo de cómo el disco es preformateado. Hay también algunos compromisos entre CLV y CAV. Aquí, las pistas son agrupadas en grupos de 20 o así, y la velocidad de rotación cambia entre las bandas, pero no dentro de las bandas. En otros discos, la velocidad de rotación es constante, pero el rango de datos es incrementado con el radio de la pista, por lo que el espaciado de bit permanece constante.

Los discos de sólo lectura (CD-ROM) y su antecesor el Compact Disc de audio usan el modo CLV. El modo CLV incluye un servo adicional encargado de controlar la velocidad del motor del disco. En este caso, la entrada es obtenida de la secuencia de datos, puesto que el código es autorreloj. Se extrae la frecuencia de los datos y se compara con la de un oscilador local y se actúa sobre el motor para que estas dos frecuencias sean iguales. La frecuencia estándar y que se

utiliza como patrón es la empleada en los primeros discos digitales de audio y es de unos 150 Kbits por segundo. Esta velocidad la podemos obtener si tenemos en cuenta que el sonido se codifica con dos canales (derecho e izquierdo) a 44.1 Kmuestras/segundo y con 16 bits (2 bytes) de resolución con lo que se obtiene $2 \times 44100 \times 16 = 1411200$, a lo que hay que añadir los códigos de detección de errores, los valores de sincronización, etc. Esta velocidad de transferencia es la imprescindible para reproducir adecuadamente un disco de audio. No obstante, los datos no están sujetos a una velocidad concreta, sino que interesa la mayor posible en cada caso. Por este motivo, han aparecido en el mercado discos que emplean una frecuencia de oscilador local que es un múltiplo de la anterior, etiquetándose comercialmente como discos X2, X8, X40, etc. según el número de veces que su oscilador es mayor que el oscilador patrón de los discos de audio. Este factor multiplicativo debe ser un número entero, para que se puedan reproducir los discos de música convencionales con un simple contador que dividirá la frecuencia por el factor apropiado, para obtener nuevamente los 150 KHz que es la única velocidad a la que deben reproducirse los discos de audio.

En la práctica, algunos dispositivos con CAV también usan servo control de la velocidad del disco para permitir al dato estar sincronizado con la señal de reloj generada dentro del dispositivo. Otros usan un motor convencional y deriva del dispositivo el reloj de una señal de reloj en el disco. Por esta razón, el preformateo puede incluir señales de reloj en dispositivos CAV y no sólo en los discos CLV.

Por regla general, los dispositivos CLV permiten aprovechar mejor la superficie completa del disco y puede proporcionar buenas velocidades de transferencias en ficheros grandes colocados de forma consecutiva sobre la pista. Por el contrario, los discos con CAV tiene unas velocidades de acceso mucho más rápidas pues no necesitan ajustar la velocidad de giro del disco. CLV resulta ideal para aplicaciones donde se precise leer grandes cantidades de datos de forma secuencial (audio, video) ya que en este caso la velocidad de giro va variando lentamente y se puede adaptar fácilmente. Por el contrario, en aplicaciones de tipo general y especialmente en bases de datos, donde hay que acceder a distintos ficheros o a cantidades de datos relativamente pequeñas pero repartidas por todo el disco, será preferible utilizar dispositivos CAV. Por este motivo, los dispositivos CD-Digital Audio, el Video-CD o el DVD emplean CLV, y los discos duros emplean CAV de forma universal.

2.20.5 Formatos de grabación

Como en los discos magnéticos, en los discos ópticos se accede a una sola pista en cada momento. El dato se escribe secuencialmente a lo largo de cada una de las pistas. Sin embargo, el factor que domina en la elección del formato de grabación óptico es el manejo de errores. El área ocupada por cada una de las celdas en el medio óptico corresponde más o menos al tamaño del punto focal, y es más pequeño que el área correspondiente del disco magnético. Típicamente, una micra de diámetro. Las técnicas de fabricación son muy diferentes; no es práctico, económicamente hablando, hacer medios que estén completamente libres de defectos y además la mayor densidad del almacenamiento óptico significa que defectos más pequeños afectarán al almacenamiento de datos, y que cada uno de los defectos afectará a más bits de información. Hay dos consecuencias: La primera es que una relativamente alta proporción de los sectores en el disco se verán afectados por los defectos, y no es tan práctico descartar todos estos sectores como haríamos con los discos magnéticos. La segunda es que la mayoría de los bits dentro del sector estarán mal con cualquier defecto, con lo que los códigos relativamente simples usados en los medios magnéticos no resultan adecuados para corregir los errores, e incluso para detectarlos todos. Un rasgo dominante del almacenamiento óptico, es la necesidad de un potente sistema de corrección de errores. No obstante los métodos de corrección de errores en sí mismos no son exclusivos de los almacenamientos ópticos sino que se emplean también en cintas magnéticas.

Hemos descrito las técnicas de grabación ópticas como aquellas en las que cada bit de dato corresponde a una celda de bit específica en el disco. De hecho, esto no es así. Como en el almacenamiento magnético, podemos usar códigos RLL y grabar grupos de códigos para incrementar la densidad de bits en el medio. Podemos usar también caracteres adicionales, o de una forma más general, usar códigos redundantes, para proporcionar detección y corrección de errores. En la grabación óptica, sin embargo, existe el riesgo de que los defectos afecten a un número bastante grande de bits sucesivos. Los métodos de codificación usados para almacenamiento óptico toman un bloque completo (o sector) de datos como una unidad, y el código es de tal forma que celdas adyacentes a lo largo de la pista correspondiente están muy dispersas dentro del bloque. Esto hace que la detección y corrección de errores sea más exacta. Los actuales algoritmos de codificación son muy complejos.

La información grabada en los discos ópticos está dividida en sectores como en los discos magnéticos, y cada sector tiene una cabecera. Estas cabeceras se ponen normalmente en el disco como parte de la información preformateada, aunque algunas controladoras usan discos en los que el preformateo se escribe por métodos termo-ópticos. En discos CAV los sectores ocupan posiciones angulares estándar, y el formato puede verse a menudo como un patrón radial. Por ejemplo en los discos magneto-ópticos, se pueden ver a simple vista los sectores y las bandas en las que se divide. En este sentido, no hay diferencia entre discos con pistas separadas y aquellos donde las pistas forman una espiral. Sin embargo, en los discos CLV, los sectores no están alineados y no pueden ser vistos por el ojo.

Las unidades de discos ópticos tienen una cabeza de lectura única y por lo tanto, el concepto de cilindro no es aplicable y las pistas son identificadas por números individuales consecutivos.

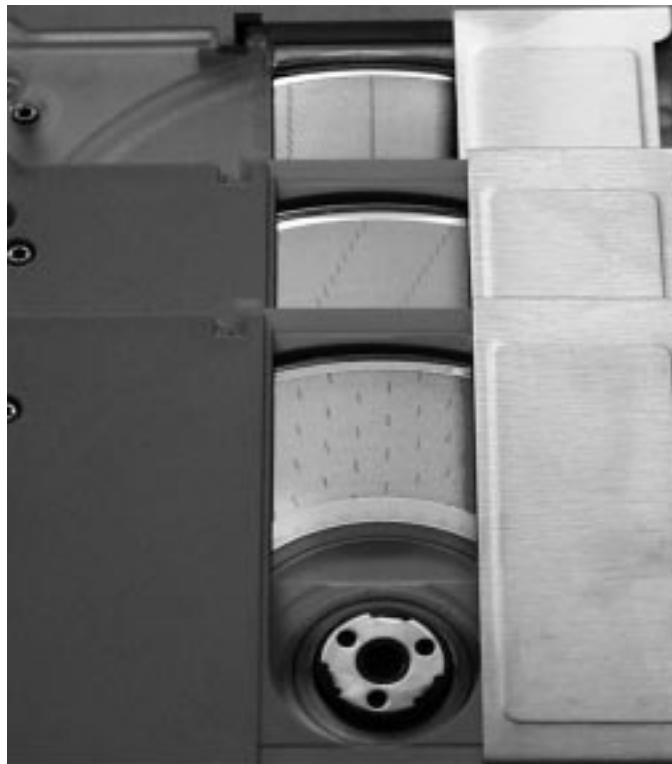


Fig. 2.29 Imagen de varios discos magneto-ópticos de 640 MB en los que se pueden apreciar los sectores y las bandas. Este último aspecto se comprueba porque las marcas de sectores no son continuas desde el centro

2.21 UN NUEVO FORMATO: EL DVD

Recientemente ha aparecido en el mercado un nuevo sistema de almacenamiento de gran capacidad basado en tecnología óptica: el 'Disco Versátil Digital' o DVD. Los principios de funcionamiento son básicamente idénticos a los de los bien conocidos discos digitales de audio (CD-DA) y CD-ROM. Sin embargo, se han introducido numerosas modificaciones que hacen que su capacidad sea sensiblemente superior.

Por una parte se emplea un láser de luz roja que tiene una longitud de onda menor (entre 635 y 650 nm) que la del CD-ROM convencional (780 nm). Esto unido a una mayor apertura focal (0.6 frente a 0.45) hace que se puedan conseguir puntos luminosos, sobre la superficie del disco, mucho más pequeños. Al disponer de una mayor focalización, con un punto de lectura ('spot') más pequeño, podemos aplicar un factor de escala a todas las dimensiones y de esta forma obtenemos una separación entre pistas de sólo 0.74 micras y una longitud mínima entre transiciones de 0.4 micras. Al tener una menor separación entre pistas, la longitud de la espiral alcanza ahora los 11 km aproximadamente, que es más del doble que en los CD's convencionales. Si ha esto se le añade que sobre esta pista los datos también están más juntos, el incremento en capacidad es considerable. La altura de los 'pits' y 'lands' también se reduce en la misma proporción por lo que en el mismo espesor (1.2 mm) de disco ahora el DVD puede contener dos capas superpuestas.

La evolución que ha sufrido la electrónica de control desde la aparición del CD de audio en 1981, permite mayores velocidades de transferencia. En el DVD, que como ya se ha anticipado emplea CLV, ésta es de 4.0 m/s frente a 1.2 m/s del CD convencional. Esto no impide que, al igual que sucedió con el CD-ROM, en el mercado haya también DVDs con velocidades x4, x6, etc. Esta evolución de la electrónica también permite la utilización de códigos de grabación y de detección y corrección de errores mucho más sofisticados que en el caso del CD-ROM. Estos códigos tienen una potencia extraordinaria y son capaces de corregir una salva de errores de hasta 2000 bytes, lo que equivale a unos 4 mm de pista. La información necesaria para la detección y corrección de errores ocupa aproximadamente un 13% de la capacidad total del disco.

La codificación EFM (Modulación de ocho a catorce) empleada en los CDs convencionales produce en algunos casos violaciones de código, para lo que se hace necesario añadir 3 bits de canal adicionales para corregirlo. Esto hace que realmente, ocho bits de datos se conviertan en la práctica en 17 bits de canal. El DVD por el contrario emplea una codificación de 8 a 16 sin violaciones de código y manteniendo o mejorando las prestaciones de la codificación empleada en los CDs a costa de una circuitería de codificación y decodificación más sofisticada. Esto proporciona un 6% de capacidad adicional.

Otra novedad que incluye el formato DVD es la utilización de ambas caras del disco e incluso permite dos capas por cada cara. Esto se consigue superponiendo una superficie semireflectante por encima de la superficie reflectante más interna. La lectura de una u otra capa se consigue focalizando el láser en distintos puntos. En base a esta característica podemos tener cuatro tipos de DVD's:

- Una cara y una capa con una capacidad de 4.7 GB
- Una cara y dos capas: 8.5 GB
- Dos caras y una capa en cada cara: 9.2 GB
- Dos caras y dos capas en cada cara: 17 GB

El diseño biestratificado de los DVDs ofrece además otra ventaja adicional: reduce los desequilibrios y el alabeo del disco. Un cambio brusco de temperatura o humedad puede alterar la planitud de un disco óptico, pero en el caso de los DVDs al tener una construcción simétrica atenúan algo este efecto que puede producir defectos de lectura.

La tabla muestra un resumen con las diferencias entre el CD convencional (CD-DA y CD-ROM) y el DVD. Las mejores características del DVD frente al CD-ROM lo hacen ideal para su utilización masiva como soporte de video, existiendo ya en el mercado un amplio catálogo de películas de cine almacenadas y disponibles en este formato. Si el CD-ROM supuso un salto considerable en cuanto al tipo de información a almacenar es de suponer que el DVD suponga una nueva revolución, permitiendo almacenar en un espacio reducido y de fácil acceso enormes cantidades de información. Aplicaciones como mapas, enciclopedias multimedia, bases de datos gráficas y sonoras, etc. pueden encontrar un aliado excelente en este tipo de soporte.

Característica	CD	DVD
Diámetro	120 mm	120 mm
Grosor	1,2 mm	1,2 mm
Estructura del disco	Sustrato único	Doble sustrato
Longitud de onda del láser	780 nm (infrarrojo)	635 - 650 nm (rojo)
Apertura numérica del sistema óptico	0,45	0,6
Separación entre pistas	1,6 μm	0,74 μm
Distancia mínima entre transiciones	0,83 μm	0,4 μm
Velocidad Lineal Constante	1,2 m/s	4,0 m/s
Número de capas	1	1 o 2
Número de caras	1	1 o 2
Capacidad	680 MB	1 capa: 4,7 GB 2 capas: 8,5 GB
Velocidad de transferencia	Modo 1: 153,6 kB/s Modo 2: 176,4 kB/s	1108 kB/s
Densidad lineal de bits	16930 bits/cm	37795 bits/cm
Densidad de almacenamiento	0,68 Gbits/cm ²	3,28 Gbits/cm ²

Tabla 2.4 Tabla comparativa entre los parámetros característicos de un CD y un DVD

Interfaces serie y paralelo

3.1 INTRODUCCIÓN

La transferencia de información entre dos sistemas digitales, por ejemplo, un microcomputador y un terminal, periférico u otro microcomputador, se realiza generalmente carácter a carácter utilizando códigos binarios (ASCII, EBCDIC, BAUDOT, CDC-Científico, ...). Otras veces la información que se transfiere no corresponde a ninguna codificación de caracteres numéricos ó alfanuméricos sino que es puramente binaria, por ejemplo, cuando se efectúan cargas de programas objeto sobre la memoria del ordenador.

De una forma o de otra la información se transmite en unidades de información denominadas palabras, que suelen ser de 8,16 o 32 bits. Existen dos formas de realizar la transmisión de estas palabras:

- ♦ Método paralelo: Transmitiendo simultáneamente, por líneas separadas, todos los bits de la palabra, junto con una señal de reloj que indica el momento en que está presente una palabra de información en las líneas de datos.
- ♦ Método serie: Transmitiendo en forma secuencial en el tiempo todos los bits de la palabra, uno tras otro, por una sola línea de datos.

Eventualmente puede existir una línea adicional de reloj que marca los tiempos de bit. El método paralelo es utilizado para transmisiones a alta velocidad entre dos sistemas; no obstante cuando la distancia entre ambos aumenta, el coste de la línea y el de los amplificadores de transmisión y recepción puede llegar a crecer de forma tal que, desde el punto de vista económico, sea preferible utilizar un sistema serie de comunicaciones. Por otra parte, los sistemas de comunicaciones serie han alcanzado un alto grado de estandarización desde hace tiempo. Existen normas universalmente aceptadas que fijan completamente todos los detalles de la comunicación, incluyendo aspectos mecánicos (tipo de conector y distribución de señales en las patillas del mismo), aspectos eléctricos (niveles y formas de las señales) y aspectos lógicos (sistemas de codificación y sincronización, y descripción de todos los circuitos de datos, control y temporizado).

Estos estándares han conducido a que la mayoría de fabricantes de procesadores y periféricos incorporen en sus equipos interfaces serie que cumplen las normas especificadas, de forma que se pueda realizar con toda facilidad la conexión indistinta de cualquier terminal o

periférico con cualquier procesador. Así, se utilizan interfaces serie para conectar periféricos, como terminales de pantalla o impresoras, a computadores aunque su distancia sea reducida y puedan, por tanto, usarse interfaces de tipo paralelo.

Las normas referidas a las interfaces paralelas, han aparecido más recientemente ya que durante más tiempo han permanecido como interfaces propietarias de los distintos fabricantes. Actualmente el nivel de normalización de éstos los sitúan a la altura de las interfaces serie.

Por último, un tercer campo en que se utilizan sistemas de manipulación de datos en serie es el de los controladores de unidades de almacenamiento de informaciones digitales sobre soportes magnéticos (discos, cassettes y diskettes). En ellos se graban y se leen los datos en forma serie, presentándose problemas de codificación comunes con los sistemas de comunicaciones serie. Recuérdese como en el tema dedicado a dispositivos de almacenamiento, los datos son almacenados en serie, por lo que aunque el dispositivo se conecte al sistema central mediante una interfaz paralela, el canal de lectura/escritura trabaja siempre en modo serie.

3.2 PROBLEMAS EN LAS TRANSMISIONES SERIE

Cuando se transmiten informaciones a través de una línea serie es necesario utilizar un sistema de codificación que permita resolver los siguientes problemas:

- a) Sincronización de bit
- b) Sincronización de carácter
- c) Sincronización de mensaje

3.2.1 Sincronización de bit

El receptor necesita saber exactamente donde empieza y donde termina cada bit en la señal recibida para efectuar el muestreo de la misma en el centro de la celda de bit. Considérese el caso de transmisión en serie de la información 01110010. Si se utilizase un método NRZ como el explicado en el tema dedicado a dispositivos de almacenamiento (no retorno a cero, en que los bits 1 ó 0 se representan por niveles 1 ó 0 respectivamente), la señal en la línea sería como la representada en la figura (3.1).

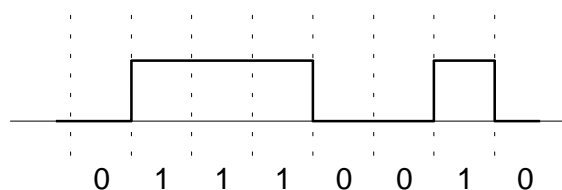


Fig. 3.1 Codificación NRZ

La presencia de varios bits iguales, por ejemplo 3 «unos», hace que la línea no efectúe ninguna transición y el receptor puede perder la referencia de donde empieza y donde acaba cada bit. Si el número de bits iguales aumenta, se observa que la dificultad de reconstruir las celdas de bit aumenta también.

Para resolver el problema de la sincronización de bit pueden usarse varios métodos:

- a) Enviar por una línea independiente de la de datos una señal de reloj que indique el centro o el inicio de las celdas de bits de la línea de datos.
- b) Enviar junto con cada bit transmitido en serie y por la misma línea una información adicional que permita al receptor extraer una señal de reloj. Esto se puede conseguir fundamentalmente

de dos formas: mediante códigos autoreloj o mediante pulsos de sincronismo como en la señal de video compuesto.

- c) Lograr, mediante algún procedimiento, que los relojes de transmisión y recepción se mantengan en fase continuamente.

Para ilustrar con mayor claridad estos conceptos veamos un proceso de transmisión y recepción con algún detalle:

- Se desea transmitir 8 bits en serie, por ejemplo 01110010.
- El transmisor, en cada flanco de subida de un reloj, envía un bit de información por la línea.
- Las señales por la línea, a lo largo del tiempo, contienen unas celdas de bit, que dependen de la frecuencia del reloj, y dentro de cada celda, el transmisor coloca un bit de información codificado según algún procedimiento, por ejemplo:
 - RZ. Una celda de bit contiene un 1 si hay un impulso positivo y un 0 si no lo hay.
 - NRZ. La celda de bit contiene un 1 (o un 0) según que el nivel de la señal sea 1 (ó 0).
 - NRZI. La celda de bit contiene un 1 si hay una transición y un 0 si no la hay.

Para que el receptor pueda interpretar adecuadamente estas señales, debe ser capaz de obtener o crear un reloj que se mantenga en perfecto sincronismo con el del transmisor. Este reloj marcará las celdas de bit y analizándolas verá si contienen un bit 1 ó 0. En este caso los datos no contienen información de reloj. Efectivamente, las secuencias de ceros, en cualquiera de los sistemas (RZ, NRZ, NRZI), y las secuencias de unos, en el sistema NRZ, no contienen ninguna transición que permita al receptor determinar la situación de las celdas de bit.

Estos sistemas se dice que no son auto-reloj y plantean el mismo problema que en los dispositivos de almacenamiento. La sincronización de bit en tales sistemas se consigue utilizando en la recepción el propio reloj de transmisión, enviado por una línea independiente de los datos, o bien utilizando relojes de precisión y con dispositivos adicionales que aseguren que se mantiene a la misma frecuencia y fase que el de transmisión. Alguno de estos sistemas de reloj se describe más adelante.

Frente a estos sistemas de codificación se encuentran los de auto-reloj (self-clock), que transmiten información de forma tal que permiten al receptor deducir la situación exacta de las celdas de bit y por tanto los datos, sin necesidad de disponer de un reloj síncrono con el de transmisión.

Hay varios métodos auto-reloj, siendo los más conocidos: PE, codificación de fase; FSK, codificación por cambio de frecuencia; FM, modulación de frecuencia; o MFM, modulación de frecuencia modificada. En estos sistemas, el envío de la información adicional para determinación del reloj se hace a costa de la disminución de la cantidad de información útil enviada para un mismo ancho de banda. Aunque como se vió en el tema de dispositivos de almacenamiento el método MFM no requiere aumentar el ancho de banda.

Dado que las características de una línea o canal de transmisión limitan la frecuencia máxima de la señal que se puede enviar por él, la cantidad que es posible enviar mediante una codificación «no auto-reloj» es normalmente mayor que mediante una codificación «auto-reloj». No obstante hay campos de aplicación idóneos para cada método.

Sin embargo, cuando el problema es de transmisión de una información serie entre dos puntos, es posible la utilización de una codificación «no auto-reloj», realizando la sincronización de bit con el propio reloj de transmisión o generando un reloj sincronizado con aquél.

Cuando el problema es de grabación de información serie en un soporte magnético giratorio (discos, cintas, etc.), para posterior reproducción o lectura, la posibilidad de utilizar el reloj usado

en la grabación, o sincronizar un reloj de recepción, se hace muy difícil al introducirse un agente perturbador como es el de las fluctuaciones de las velocidades de giro del soporte magnético en los instantes de grabación y lectura.

En tales casos es preferible utilizar un método auto-reloj y disponer en el receptor un circuito que extraiga el reloj de recepción de los datos, haciéndose insensible a posibles derivas del reloj de grabación y de las velocidades del soporte.

3.2.2 Sincronización de carácter

La información en serie se transmite, por definición, bit a bit, pero la misma tiene sentido en palabras, por ejemplo de 8 bits. El sistema de codificación usado debe permitir distinguir sin ambigüedades dentro de una corriente de bits cuáles son los 8 bits que forman una palabra. Normalmente se resuelve enviando los bits de cada carácter separados por alguna señal de sincronismo.

Por ejemplo, la siguiente información en serie: 0100110001001100100
Puede tener distintas interpretaciones según como se agrupen los 8 bits para formar las palabras. 01, 00110001, 00110010, 0 o también 010, 01100010, 01100100.

La primera agrupación representa los caracteres 1 y 2 según una codificación ASCII, la segunda representa, según la misma codificación, los caracteres b y d.

Para obtener la sincronización de carácter pueden utilizarse diversos sistemas, unos se basan en la utilización de líneas adicionales a las de datos para enviar impulsos que indican el inicio de un bloque de caracteres. Tal impulso identifica el primer bit del primer carácter de un bloque o mensaje, y luego, por conteo de bits y caracteres se determinan todas las fronteras de los datos del bloque.

Otros sistemas, utilizados usualmente en los sistemas de comunicaciones serie, son:

Asíncrono: Cada carácter va señalizado mediante dos bits, uno al principio, bit de arranque, y otro al final, bit de parada. Estos bits permiten reconocer las fronteras de los caracteres.

Síncrono: Cada mensaje o bloque de transmisión va precedido por unos caracteres de sincronismo. Cuando el receptor identifica una configuración de bits igual a la de los caracteres de sincronismo, da por detectado el inicio de los datos y a continuación, por conteo de bits y caracteres identifica todos los caracteres del bloque.

3.2.3 Sincronización de mensaje

En un sistema de comunicaciones generalmente las informaciones se transmiten en bloques de caracteres. Por sincronización de mensaje entendemos el mecanismo por el cual un conjunto de palabras es interpretado correctamente. Este problema normalmente no incumbe a los circuitos de codificación, sino al procesador que lo utiliza. El conjunto de reglas (protocolo) que permiten interpretar correctamente los mensajes suele estar controlado por una tarea software (un programa) que ejecuta el ordenador, aunque actualmente hay ciertos circuitos integrados LSI que efectúan alguna de estas tareas.

3.3 MÉTODOS DE E/S PARA COMUNICACIONES SERIE

En lo que sigue se utiliza el nombre genérico de terminales para designar a los sistemas que se comunican utilizando un procedimiento serie de entrada/salida. Un terminal puede ser un ordenador, un microcomputador, un periférico, etc. La comunicación entre terminales se hace utilizando líneas o canales de transmisión, que pueden ser:

Simplex: cuando son capaces de transmitir información en un solo sentido.

Semiduplex (half-duplex): cuando son capaces de transmitir información en ambos sentidos pero no de forma simultánea.

Dúplex (full-duplex): cuando son capaces de transmitir simultáneamente información en ambos sentidos.

La codificación de las señales en estos sistemas se hace mediante uno de los siguientes métodos: asíncrono o síncrono.

3.3.1 Método asíncrono

En el método asíncrono la transmisión se controla por bits de inicio y de final que enmarcan cada carácter transmitido, son los denominados bits de inicio ('start') y parada ('stop') y son utilizados por el terminal receptor para sincronizar su reloj con el del transmisor en cada carácter. La especificación RS404 de EIA (Electronic Industries Association) define las características del método asíncrono de transmisión serie. La transmisión en asíncrono se basa en las siguientes reglas:

- a) Cuando no se envían datos por la línea, ésta se mantiene en estado 1.
- b) Cuando se desea transmitir un carácter se envía primero un bit de inicio, que pone la línea a cero durante el tiempo de 1 bit.
- c) A continuación se envían todos los bits del carácter a transmitir con los intervalos que marca el reloj de transmisión.
- d) A continuación del último bit del carácter se envía el bit de final que hace que la línea se ponga a 1 por lo menos durante el tiempo de 1 bit.

Los datos codificados según estas reglas pueden ser detectados fácilmente por el receptor. Para ello deben seguirse los siguientes pasos:

- 1) Esperar una transición de 1 a 0 en la señal recibida.
- 2) Activar un reloj de frecuencia igual a la del transmisor.
- 3) Muestrear la señal recibida al ritmo de este reloj para formar el carácter.
- 4) Leer un bit más de la línea y comprobar si es 1 para confirmar que no ha habido error de sincronización.

El bit de final tiene la misión de llevar la línea a estado 1 para que el bit de inicio del próximo carácter provoque la transición de 1 a 0 que permita al receptor sincronizar el siguiente carácter. El bit de final sirve también para dar tiempo a que el sistema receptor acepte el dato recibido. De todas formas, actualmente se utiliza siempre una memoria de tipo FIFO que almacena el dato recibido mientras el receptor está recibiendo el siguiente, de forma que el procesador dispone del tiempo de todo un carácter para recogerlo.

El método asíncrono de transmisión presenta las siguientes ventajas:

- 1) Permite enviar caracteres a ritmos variables ya que cada uno de ellos lleva incorporada la información de sincronismo.
- 2) Existen circuitos integrados de bajo costo, las UART (Universal Asynchronous Receiver Transmitter), que simplifican enormemente la realización de sistemas de entrada/salida en este formato.
- 3) Es un método de comunicaciones estándar entre ordenadores y terminales de pantalla, impresoras lentas, ratones, modems, etc.

Entre sus inconvenientes se puede citar, como más importante, su ineficiencia, ya que cada carácter va lastrado con dos bits de sincronización que no contienen información útil. Suponiendo caracteres de 8 bits, es necesario enviar por la línea 10 bits para enviar un carácter, es decir sólo un 80% de la información transmitida es válida.

3.3.2 Método síncrono

En el método síncrono, en vez de añadirse bits de sincronismo a cada palabra, lo que se hace es añadir caracteres de sincronismo a cada bloque de datos. Los caracteres se transmiten en serie, bit a bit, y sin ninguna separación entre uno y otro, no obstante, delante de cada bloque de datos se colocan unos caracteres de sincronismo que sirven al receptor para realizar la sincronización de carácter, es decir, conocer las fronteras de carácter en una corriente de bits. La sincronización de bit se consigue normalmente utilizando una señal externa de reloj. En una comunicación local entre dos dispositivos, el transmisor envía por una línea independiente de la de datos su señal de reloj, que es utilizada por el receptor como reloj de recepción. La sincronización de bit queda de esta forma resuelta, ya que el mismo reloj que el transmisor utiliza para serializar los bits de información sobre la línea de datos, es utilizada por el receptor para leer los datos recibidos. Será necesario únicamente tener en cuenta que el receptor debe muestrear la línea de datos con el flanco de reloj contrario al que el transmisor utilizó para enviarlos, para que así el muestreo se efectúe en el centro de la celda de bit.

El método de comunicaciones síncrono se utiliza cuando el volumen de información a enviar es importante, debido a su mayor eficiencia respecto al método asíncrono. Como ya se ha comentado, en modo asíncrono cada palabra se envía precedida por un bit de inicio y seguida por 1 ó 2 bits de final. Suponiendo palabras de 8 bits y utilización de 1 bit de final, se necesitan 10 bits para enviar una palabra de 8 bits. En modo síncrono, cada mensaje se envía precedido por unos caracteres de sincronismo, normalmente dos caracteres SYN (ASCII N° 22).

Para enviar un mensaje de N palabras serán necesarios $(N + 2) \times 8$ bits en síncrono y $10 \times N$ bits en asíncrono. Comparando ambas cifras se observa que para mensajes superiores a 8 bits el sistema síncrono es más eficiente, y para mensajes de 512 octetos la eficiencia del método síncrono es un 25 % superior a la del método asíncrono.

3.3.3 Regeneración del reloj en el receptor

En una comunicación remota utilizando modems, la señal de reloj es extraída del canal de datos por el modem; para ello utiliza un reloj de la misma frecuencia que el transmisor y que mediante circuitos de sincronización lo mantiene en la misma fase. El sistema es inherente al principio de funcionamiento del módem. Existen los llamados modems síncronos y modems asíncronos. Los modems asíncronos utilizan sistemas de codificación FSK cuya misión es generar una señal de distinta frecuencia para la marca "1" y el espacio "0", esta señal debe ser de frecuencia apropiada para que pueda transmitirse a través de la red telefónica, ya que ésta sólo permite la banda de audiofrecuencia. El módem receptor recibe la señal de la línea telefónica y discrimina los dos tonos generando las señales marca y espacio que reconstruyen la señal digital primitiva. Debido a este modo de funcionamiento, el módem en sí no está ligado a la frecuencia de

transmisión de los datos y admite, sin necesidad de ningún ajuste, señales de frecuencias de transmisión comprendidas entre cero y el máximo.

Otra forma de sincronizar los dos sistemas es el emplear en el receptor un oscilador de una frecuencia varias veces superior (normalmente $\times 16$) y que es el método que se emplea por ejemplo en los circuitos que incorporan los PC's para los puertos serie. Según este método, el receptor genera una señal de frecuencia 16 veces superior a la empleada por el emisor para enviar los datos y toma la muestra en el centro, es decir cuando la señal de recepción haya completado 8 ciclos, tal y como se muestra en la figura (3.2). Los flancos de la señal recibida se emplean para resincronizar el oscilador local de recepción. Esto es necesario, porque no se puede garantizar que el oscilador de recepción sea exactamente 16 veces el de emisión. Si existe alguna pequeña deriva, y por ejemplo el oscilador de recepción va ligeramente más lento de lo que debería, el siguiente flanco de señal llega antes de que complete la cuenta de 16 ciclos. Si esto sucede, la cuenta vuelve a comenzar desde cero (aunque no hubiese completado la anterior). Esto se puede conseguir con el circuito que se muestra en la figura (3.3) donde se utiliza un contador que reinicia la cuenta cada vez que llega un flanco en la señal de datos.

El objetivo es conseguir que el reloj del receptor esté en fase lo más exactamente posible con el reloj del transmisor y para ello se aprovecha cada flanco de la señal recibida para reiniciar el ciclo del reloj de recepción. La salida de peso más alto del contador de 4 bits es el reloj de frecuencia f . Esta señal tiene el flanco de subida en el momento en que el contador se pone a 0. La sincronización se efectúa mediante la puesta a 0 asíncrona del contador cada vez que se detecta un flanco en la línea de datos. En el momento del flanco, el contador debe estar a 0; si no lo está, significa que el reloj tiende a adelantarse o atrasarse, y se aprovecha este momento para ponerlo en sincronismo.

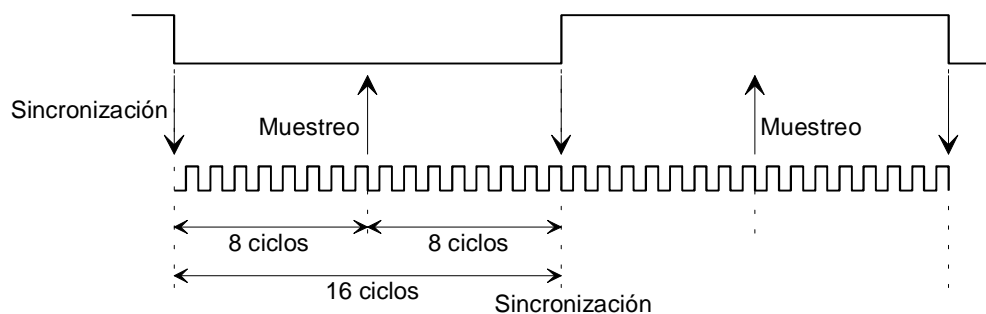


Fig. 3.2 Método de sincronizar emisor y receptor con un reloj de recepción 16 veces más rápido que el de emisión.

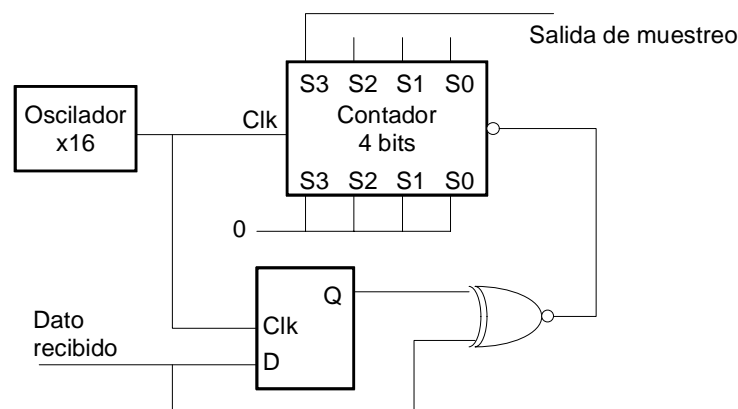


Fig. 3.3 Circuito de sincronización del reloj de recepción.

El sistema de detección de flancos utiliza una báscula D para generar una salida retrasada 1/16 de bit respecto a la señal de entrada, la puerta NOR exclusiva genera un impulso de 0 en los instantes en que la señal D varía.

Otro sistema es el basado en un lazo de sincronización de fase digital (DPLL: Digital Phase Locked Loop). Por ejemplo, el 8273, controlador de comunicaciones para protocolos tipo HDLC y SDIC, incorpora un PLL digital que funciona de la siguiente forma:

Se utiliza un reloj de frecuencia 32 veces superior a la de transmisión. A partir de este reloj $32 \times R$ y del reloj recibido, el DPLL genera un impulso que está centrado en las celdas de bit. El DPLL reacciona contra derivas y distorsiones de fase en los datos recibidos haciendo correcciones en la fase del impulso de reloj a base de incrementos discretos. El impulso del DPLL se genera a partir de un conteo de 32 impulsos del reloj $32 \times R$, con una corrección de -2, -1, +1 ó +2 según el cuadrante en que se detecta el flanco de la señal recibida. En el ejemplo de la figura (3.4) el flanco de la señal se detecta en el 2º cuadrante indicando que el impulso no estaba en el centro de la celda de bit sino marcadamente desfasado hacia la derecha; por tanto el sistema correctivo actúa haciendo que el próximo impulso en vez de salir 32 tiempos de $32 \times R$ más tarde, salga sólo 31 tiempos más tarde, iniciándose el proceso de centraje del impulso en el centro de la celda de bit. Puede demostrarse que, partiendo de una señal recibida en reposo, el DPLL tarda 12 tiempos de bit en sincronizarse en el peor caso. Para ello es necesario enviar antes de cada bloque unos caracteres que posean suficiente número de flancos para que el DPLL pueda conseguir la sincronización de bit de forma rápida.

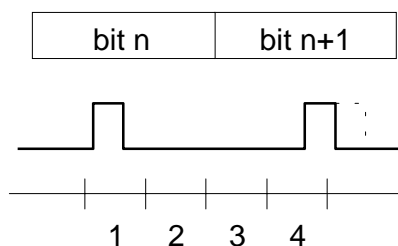


Fig. 3.4 Resincronización utilizando un PLL digital

3.4 ESTÁNDAR DE COMUNICACIÓN SERIE RS-232

Los sistemas de comunicaciones serie tienen a su disposición un conjunto de recomendaciones elaboradas por asociaciones e institutos de normalización (ISO, EIA, CCITT, etc.) que especifican con precisión todas las características del sistema de comunicaciones. Las normas para comunicaciones serie están clasificadas por niveles; aquí interesa resaltar únicamente el NIVEL 1, que hace referencia a:

- a) Las características eléctricas de las señales
- b) Las características mecánicas de la interfaz
- c) La descripción funcional de las señales.

La norma más ampliamente aceptada es la EIA RS-232-C, que define las características funcionales, eléctricas y mecánicas de la interfaz entre un ordenador o terminal y un equipo de comunicaciones (por ejemplo un módem). La norma RS-232-C puede ser aplicada a la conexión entre dos ordenadores, aunque no se utilicen modems, como se verá a continuación. Las especificaciones funcionales de la RS-232-C coinciden con la recomendación V.24 del CCITT

(Comité Consultatif International Telephonique et Telegraphique) y definen 21 circuitos con el significado que se muestra en la tabla (3.1). (Se utiliza la numeración de circuitos según CCITT):

Estas son 21 señales que RS-232-C y V.24 especifican para la comunicación entre un terminal y un módem. Para la comunicación entre dos terminales, sin utilización de modems, se utiliza un subconjunto de 3, 5 ó 7 señales solamente, aunque se respetan sus especificaciones funcionales, eléctricas y mecánicas. A este tipo de conexión se le denomina modem nulo para resaltar la ausencia de éste. En cuanto a especificaciones mecánicas, la norma RS-232-C establece un conector de 25 patillas y fija todas sus dimensiones, así como la distribución sobre el mismo de las 21 señales (Tabla 3.1) Algunos sistemas utilizan un subconjunto de estas señales y emplean un conector de 9 contactos.

Pin	Abrev.	Función
1		Tierra de protección
2	TxD	Transmisión desde el terminal
3	RxD	Recepción desde el módem
4	RTS	Petición de envío
5	CTS	Listo para enviar
6	DSR	Dato preparado
7		Masa de señal común
8	DCD	Detector de portadora
9		Reservado
10		Reservado
11		Sin asignar
12		Detector secundario de portadora
13		Listo para enviar secundario
14		Transmisión de datos secundario
15		Reloj de transmisión desde el módem
16		Recepción de datos secundario
17		Reloj de recepción
18		Sin asignar
19		Petición de envío secundario
20	DTR	Terminal de datos preparado
21		Detector de calidad de señal
22	RI	Indicador de llamada (Timbre)
23	DSRD	Selector de velocidad de la señal de datos
24		Reloj de transmisión desde el terminal
25		Sin asignar

Tabla 3.1 Circuitos especificados por CCITT en la recomendación V.24

Uno de los elementos principales que hay que fijar para cualquier comunicación son los niveles eléctricos de la señal, si emplea lógica positiva o negativa, si la transferencia es por niveles de tensión o de corriente, si es en modo diferencial o no, etc. En las comunicaciones serie, los niveles más habituales son los TTL y los RS-232.

Señales TTL: Los niveles de tensión más inmediatos para la transmisión de señales son los correspondientes a las señales TTL ya que son los que habitualmente emplean internamente los distintos sistemas. Sin embargo no son adecuados para transmisión de datos por diversos motivos y deben emplearse otras soluciones alternativas. La comunicación basada en señales TTL está basada en el envío directo por una línea unifilar o por pares trenzados de las señales de salida de las puertas TTL. No es aconsejable su utilización para distancias mayores de 5 metros lo cual restringe considerablemente su rango de aplicación. La figura (3.5) muestra los niveles de tensión

correspondientes a estas señales. Una puerta TTL envía un '1' lógico poniendo en su salida una tensión entre 2.4 y 5 Voltios, y envía un '0' lógico poniendo en su salida una tensión inferior a 0.4 Voltios. Por otra parte, una puerta TTL interpreta la entrada como '1' lógico cuando la tensión de entrada es superior a 2 Voltios y un '0' lógico cuando la entrada es inferior a 0.8 Voltios.

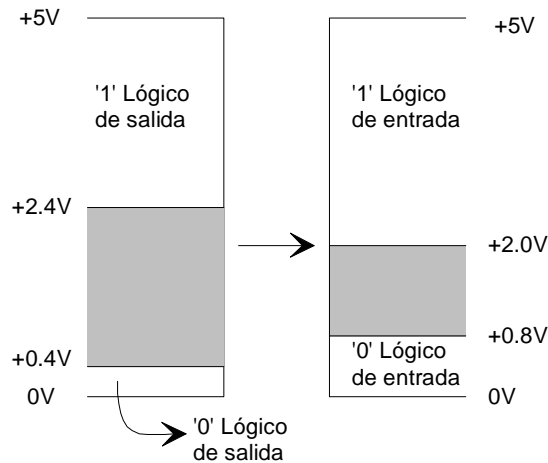


Fig. 3.5 Niveles de tensión correspondientes al '0' y al '1' en la salida y en la entrada de una puerta TTL

Señales RS-232: La figura (3.6) muestra los niveles eléctricos estándar de la RS-232 que emplea lógica negativa. Un '1' lógico corresponde a valores de tensión entre -5 y -15 V, es decir 'LO'. Un '0' lógico corresponde a valores de tensión entre +5 y +15 V, es decir 'HI'. Estos niveles son para circuitos cargados, en vacío los niveles pueden variar entre ± 25 V. El receptor admite rangos de +3 a +25 V para el '0' lógico y de -3 a -25 V para el '1'. La amplia región entre ± 3 V minimiza los problemas de ruido y permite una operación fiable hasta 15 metros de distancia.

Bucle de corriente: Permite realizar comunicaciones a mayores distancias, hasta 300 metros según la velocidad (normalmente 1200 bps a 30 y 10 bps a 300 m). Los niveles 1 y 0 se codifican por la ausencia o presencia de una corriente unidireccional de 20 mA en la línea.

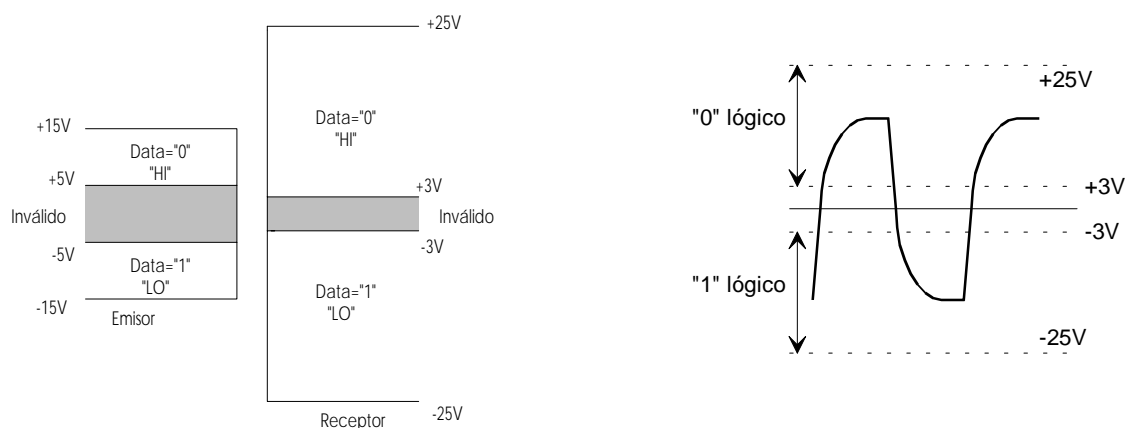


Fig. 3.6 Niveles de tensión en el emisor y receptor según la norma RS-232

Utiliza un conector estándar de 25 terminales estando adoptado por todos los fabricantes el tipo DB-25 con la asignación de pines que se muestra en la tabla (3.1). El módem (equipo de comunicación de datos o DCE) incorpora un conector hembra, mientras que el terminal o DTE (normalmente un ordenador) dispone de un conector macho. De las 21 señales definidas,

generalmente sólo se utilizan nueve y a menudo sólo son tres las empleadas: emisión, recepción y masa de señal que se corresponden con los terminales 2, 3 y 7 del conector de 25 terminales.

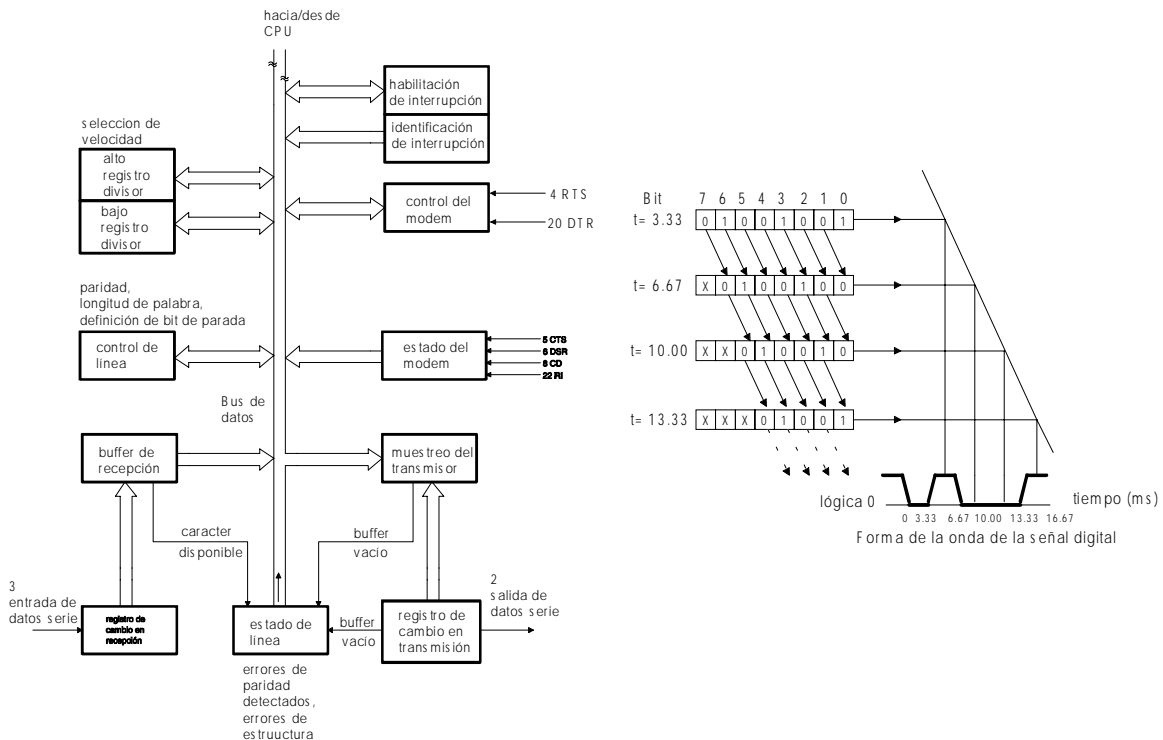


Fig. 3.7 Esquema de funcionamiento del integrado NS8250

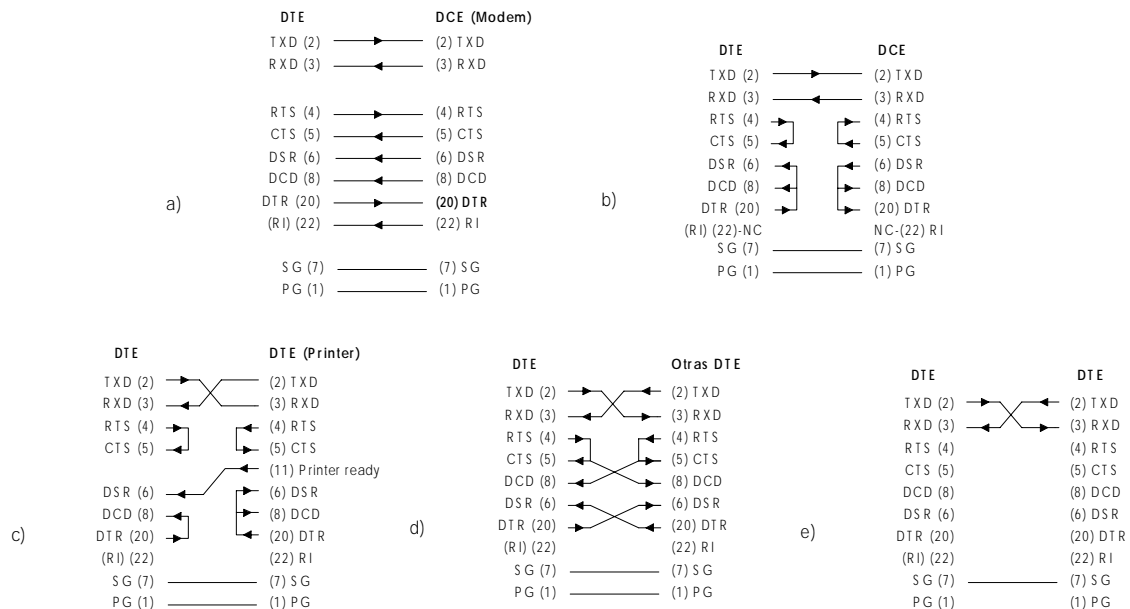


Fig. 3.8 Distintas configuraciones de conexión con RS-232

El tipo de circuitos que realizan este tipo de comunicaciones reciben el nombre de UART (Universal Asynchronous Receiver Transmitter). El interfaz RS-232C es implementado en el adaptador asíncrono de comunicaciones del PC, usando el chip de National Semiconductor INS8250 UART, cuyo esquema de funcionamiento se muestra en la figura (3.7). Este adaptador es capaz de comunicar a diferentes velocidades desde 50 a 9600 baudios. El usuario puede controlar la velocidad de transmisión, la longitud de caracteres, paridad y bits de parada a través del sistema

operativo o programas de usuario. Versiones más recientes del integrado permiten velocidades muy superiores. Por ejemplo el integrado 16550 alcanza los 115200 bps e incluye una memoria FIFO de 16 caracteres para emisión y otra de idéntica longitud para recepción. En la figura (3.8) se muestran distintas configuraciones de conexión con RS-232.

3.4.1 Variantes RS-422, 423 y 485

RS-232-C utiliza emisores y receptores no balanceados, la señal '1' es una tensión $\leq -3V$ y la señal '0' es una tensión $\geq +3V$. Se utiliza normalmente una señal de +12 y -12 V (la especificación indica $\pm 3V$ a $\pm 25 V$.) La velocidad de subida de la señal se limita a $30V/\mu s$. esta interfaz está especificada para una velocidad máxima de transmisión de 20 kbps y una distancia de 15 m.

Cuando se requieren velocidades mayores de transmisión que las que ofrecen los anteriores sistemas es necesario utilizar un sistema de transmisión diferencial, para evitar los efectos del ruido que aparecen con tensiones en modo común en las salidas del emisor o a la entrada del receptor. La norma RS-422 fue definida por la EIA para este propósito permitiendo velocidades de transmisión de hasta 10Mbits/s y hasta una longitud de 1200m. Los dispositivos emisores que cumplen esta norma son capaces de transmitir señales diferenciales con un mínimo de 2V de diferencia sobre un par de líneas trenzadas y terminadas con una impedancia de 100 Ohms. Los receptores deben ser capaces de detectar una señal diferencial de $\pm 200mV$. en presencia de una señal común de $\pm 7V$. Se utilizan señales de hasta 6 V y el receptor tiene un umbral de disparo de 200mV.

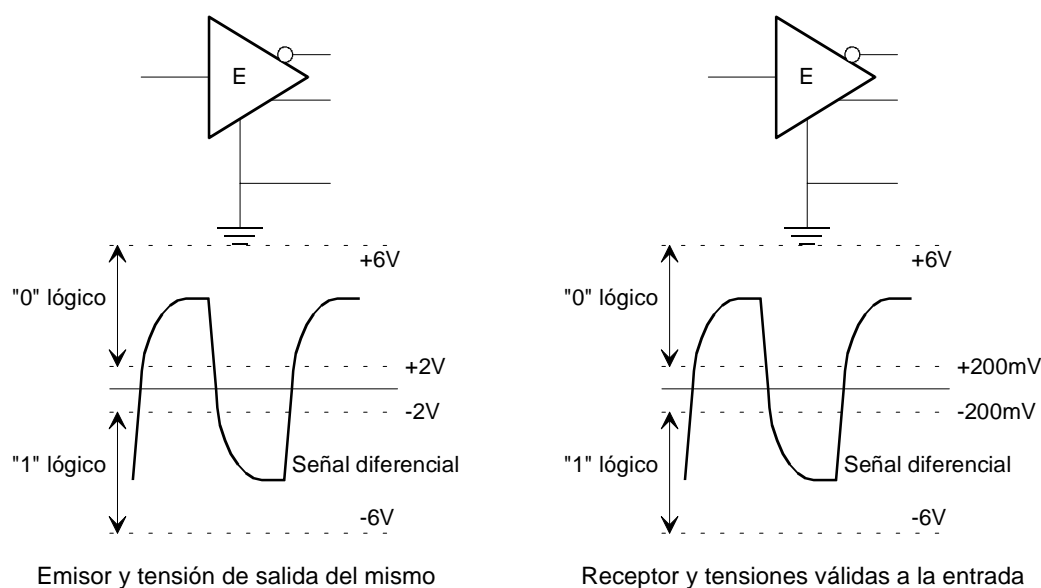


Fig. 3.9 Niveles de tensión de salida y de entrada de los interfaces RS-422 y RS-485

Una ventaja de esta norma frente a RS-232-C es que en aplicaciones de bus, permite que un solo emisor pueda comunicar con varios receptores aunque tiene la limitación de que los restantes receptores deben tener una alta impedancia de entrada para no cargar el bus. Un problema que presentan ambos interfaces es el de la contención. Es decir, no permite que varios emisores transmitan información simultáneamente. Cuando esto ocurre, la excesiva corriente producida por la tensión en modo común generada, puede llevar a la destrucción del circuito emisor, puesto que no existen limitaciones para evitarla.

Una situación intermedia entre las dos normas comentadas es la propuesta en la norma RS-423. Ésta utiliza un receptor diferencial y un emisor que no lo es; de esta forma se permite su

interconexión con emisores o receptores RS-232C y RS-422 indistintamente. Las prestaciones que se consiguen son: 300Kbps a 12 m y 3Kbps a 1200 m.

La principal ventaja de las normas 422 y 423 es que utilizan recepción en par diferencial o transmisión balanceada, lo que las hace más inmunes al ruido. Esto es debido a que las variaciones introducidas por el ruido debido a interferencias electromagnéticas, afectarán por igual a las dos señales, y como el receptor toma el dato de la diferencia entre ambas, no se verá afectado por esta situación. Lo único que se requiere es que el receptor tenga un alto factor de rechazo al modo común, es decir que sea insensible a las variaciones conjuntas de sus dos terminales de entrada.

Para solventar algunos problemas que presentaban las anteriores normas, la EIA definió un nuevo estándar: la norma RS-485. Se considera como una interfaz multipunto (ver figura 3.10) y permite la comunicación de hasta 32 pares de emisores-receptores en un bus de datos común satisfaciendo al mismo tiempo los requerimientos de la RS-422. Las diferencias fundamentales son las siguientes:

- Margen de tensiones ampliado hasta -7V a +12V frente a -0.25 a +7 de la RS-422
- El emisor dispone de protección frente al problema de la contención.
- El margen de tensiones en el receptor va desde $\pm 7V$ a $\pm 12V$ manteniendo una sensibilidad de $\pm 200mV$.
- Incremento de la impedancia de entrada del receptor hasta 12 Kohms.

Para concluir este apartado hay que resaltar que las diferencias entre las distintas variantes, están únicamente en la capa física. Es decir los circuitos de tipo UART pueden ser idénticos, con lo que la programación es independiente de la variante que se esté utilizando. El emplear una u otra depende de los circuitos excitadores de línea y de los correspondientes circuitos receptores. La tabla (3.2) resume las características de las distintas variantes.

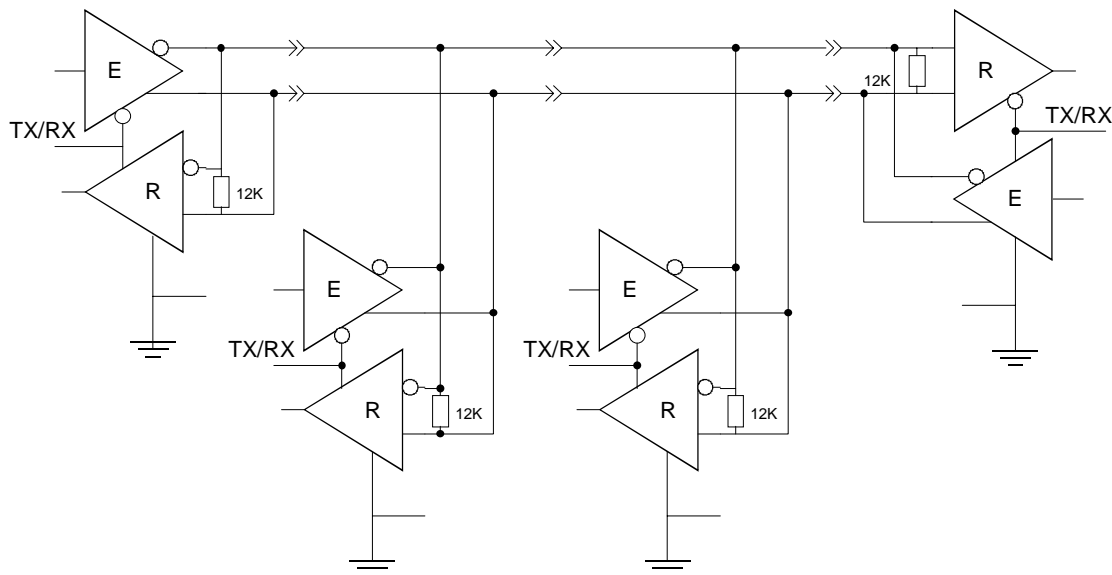


Fig. 3.10 Configuración típica de RS-485 mostrando 4 equipos conectados simultáneamente al mismo bus

Parámetro	RS-232	RS-422	RS-423	RS-485
Modo de trabajo	Simple	Diferencial	Simple	Diferencial
Nº de emisores permitidos	1	1	1	32
Nº de receptores permitidos	1	10	10	32
Longitud máxima del cable	15m	1200m	1200m	1200m
Vel. de transmisión (Bps)	20K	100K	10M	10M
Tensión en modo común	±25	+6V -0.25	±6	+12 -7
Tensión de salida	±5 V mín ±15 V máx	±2 V	±3,6 V mín ±6.0 V máx	±1.5 V mín
Carga de excitación	3kΩ-7kΩ	100Ω	450Ω mín	60Ω
Pendiente de subida	30V/μs máx	X	Determinado por la longitud del cable	X
Impedancia de entrada	3kΩ-7kΩ	4kΩ	4kΩ	12kΩ
Sensibilidad del receptor	±3 V	±200 mV sobre un modo común de ±7 V	±200 mV	±200 mV sobre un modo común de ±12 V

Tabla 3.2 Resumen de características de las distintas variantes de la interfaz RS-232

3.5 EL INTERFAZ MIDI

3.5.1 Un poco de historia

El interfaz MIDI fué diseñado originalmente para la interconexión de instrumentos musicales digitales entre sí y de estos con un ordenador. No es de extrañar por tanto que su gestación y sus características estén estrechamente relacionadas con el mundo de los instrumentos musicales y de la música y el espectáculo en general.

A comienzos de los setenta comenzaron a aparecer los primeros sintetizadores electrónicos de tipo analógico. Un sintetizador es un instrumento que genera los sonidos musicales a partir de elementos electrónicos básicos, como osciladores, generadores de envolvente o de rampa, filtros, etc. Este primer tipo de sintetizadores se podían conectar entre sí de forma analógica con una señal que proporcionaba una tensión de 1 voltio por octava. De esta forma, señales que se diferenciaban entre sí en un voltio, representaban la misma nota pero de octavas adyacentes. Con este sencillo interfaz electrónico, se podía gobernar un sintetizador de un fabricante con un teclado de otro.

Rápidamente se comenzó a introducir la digitalización y el control por ordenador de los múltiples osciladores de los sintetizadores polifónicos. A partir de este momento, el sencillo interfaz analógico de un voltio por octava dejó de ser aplicable y los instrumentos volvieron a estar incomunicados entre sí. Algunas compañías, a la vista de las limitaciones impuestas a los usuarios, comenzaron a desarrollar estructuras de bus capaces de permitir distintas expansiones. Algunas de ellas empleaban el sistema de bus serie, con objeto de rebajar el coste, mientras otras elegían el bus paralelo debido a su mayor rapidez. En lo único que estaban todas de acuerdo era en la necesidad de desarrollar una interfaz apropiada y común a todas ellas.

En diciembre de 1982, Sequential Circuits Inc. (fabricante del Prophet, primer sintetizador polifónico de difusión masiva) lanzó las primeras unidades del Prophet 600. Una de sus

características más interesantes era que disponía de una conexión de interfaz serie, y que Dave Smith, presidente de Sequential denominó entonces como Universal Synthesizer Interface (USI). Durante la Feria de Invierno de Fabricantes de Música que se desarrolló aquel mismo año, técnicos de Sequential, Yamaha y algunos otros fabricantes celebraron una reunión informal en la que comenzaron a discutir las bases de una posible estandarización. Como resultado de estas conversaciones apareció un protocolo muy similar al USI de Dave Smith, y que ofrecía la mejor relación entre velocidad, simplicidad y bajo costo.

El junio de 1983 se conectó un Prophet 6000 a un Yamaha DX-7 (instrumento basado en el afortunado, y no por ello menos importante descubrimiento de John Chowning: la técnica de sintetizado en FM. Técnica que modificaría el mundo de los teclados para siempre). El resultado careció por sí mismo de espectacularidad pero motivó el que en agosto de 1983, representantes de Sequential, Roland, Yamaha, Korg y Kawai sentaran en Tokio las bases de la norma «MIDI 1.0» (Musical Instrument Digital Interface).

Esta interfaz es quizá la mas extendida y prácticamente única dentro de su campo de aplicación pese a no haber sido respaldada por ningún organismo internacional de normalización, pero no hay instrumento musical electrónico que se precie, que no disponga de una conexión MIDI, e incluso resulta muy económico incorporar este tipo de interfaz a cualquier ordenador. De hecho, casi cualquier tarjeta de sonido convencional incluye uno.

3.5.2 El hardware MIDI

La característica técnica más importante del MIDI reside en que, para abaratar los cables y las conexiones, se ha usado un protocolo serie, básicamente el mismo que el RS-232 con un bit de comienzo (START), 8 bits de datos y dos bits de fin (STOP). Funciona a una velocidad de 31,25 kilobaudios, lo que a primera vista sin duda parece un poco extraño. Pero 31,25 Kbd. no es una velocidad tan extraña si consideramos el aún popular (y bastante barato hoy día) adaptador de comunicaciones asincrónicas 6850, cuyos registros de control internos actúan en modo división por 64, y son controlados externamente por un reloj de transmisión/recepción que funciona a 2 MHz., lo que arroja un resultado de 31,25 Kbd. A diferencia del RS-232 (que usa una tensión bipolar), el MIDI utiliza un bucle de corriente de 1,5 mA con un optoacoplador en la entrada del receptor (ver figura 3.11). En esta figura se muestra la entrada MIDI, la salida MIDI y una salida especial denominada MIDI Thru. Si observamos la figura, observamos que esta salida no es más que una réplica de la entrada convenientemente optoaislada de esta. Si no existiera esta salida, tan sólo se podrían conectar dos equipos. Como los datos emitidos por este tercer puerto son una réplica de los recibidos por el dispositivo en MIDI IN, su uso permite el encadenamiento de varios dispositivos. Aunque en teoría, la interconexión via MIDI Thru es transparente, en la práctica se produce una distorsión que puede acarrear la pérdida de mensajes después de más de tres enlaces. Por este motivo, en entornos complejos con múltiples dispositivos, es aconsejable emplear un dispositivo especial denominado expansor MIDI o "MIDI patch bay".

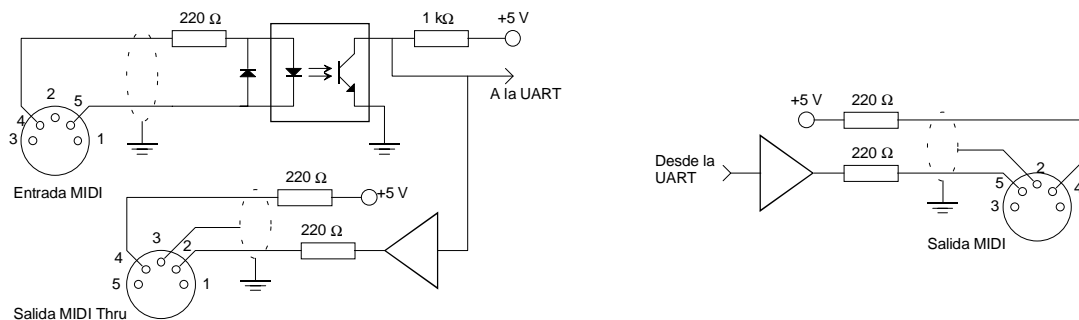


Fig. 3.11 Entradas y salidas en un interfaz MIDI

En cuanto a los conectores, el MIDI usa los del tipo DIN de 5 patillas, con hembras en los equipos y machos en los cables. Este tipo de conectores son perfectamente válidos para usarse tanto en equipos de consumo como de estudio. Tampoco pasa desapercibido el hecho de que, por otra parte, los conectores pueden siempre ser sustituidos a la hora de la verdad por los de clase XLR profesional.

El cable que se usa en los equipos de MIDI suele ser un par trenzado y blindado cuya longitud no exceda de 15 metros. Como sólo necesita dos hilos y la pantalla, resultan conexiones económicas. Sólo hay una precaución que tomar: hay en el mercado cables terminados en conectores de tipo DIN de 5 patillas y destinados a equipos de audio. Debido a su bajo precio podemos caer en la tentación de usarlos como sucedáneo de los cables auténticos de MIDI. En la mayoría de los casos no tendremos problemas, pero el riesgo existe. ¿Por qué? Si un cable no es más que un conjunto de hilos. Pues NO, no siempre es así. El problema surge debido a dos causas distintas, cada una de ellas insignificante por sí sola. Debido a que la mayoría de los conectores DIN van soldados sobre una placa tiene sentido que éstos vayan soldados a masa. En algunos equipos la masa es también la tierra del sistema (lo cual parece apropiado), y en la mayoría de las tomas DIN la carcasa va conectada electrónicamente al blindaje. El que estos cables de audio lleven conectados a una de sus patillas no es problema, pero sí lo es el hecho de que sus blindajes estén interconectados a través del apantallado del cable. Si uno de los cables se usa para conectar dos partes de un equipo cuyos diseñadores han puesto a tierra las carcasas de los conectores, el resultado es una realimentación instantánea, y no se trata sólo de un posible zumbido de audio; sino que éste está casi garantizado. Esto se producirá por lo que se conoce como lazos de tierra que consiste en que a través de los blindajes de los cables que interconectan los distintos equipos pasa una corriente nada despreciable consecuencia de que no todos los equipos tienen el mismo potencial de tierra.

Centrándonos en la figura (3.11), y recordando que una de las premisas fundamentales del sistema MIDI es prevenir dichas realimentaciones antes de que ocurran, observemos que ninguno de los pines del conector MIDI-IN está conectado a masa, mientras que en el conector MIDI-OUT sólo el pin 2 va a masa. Esta es una característica particularmente importante en los sistemas MIDI: no hay masa común entre equipos a través del cable de conexión.

A pesar de que la figura (3.11) muestra tres tipos de conectores MIDI (IN, OUT y THRU), la mayoría de equipos sólo incorporan los de tipo IN y OUT. Algunos módulos de expansión de voces sin teclado sólo montan el IN, mientras que otros como los controladores de teclado o las bases de tiempo pueden llevar sólo el de tipo OUT. Otros equipos de mayores prestaciones incorporan el conector THRU. Dicho conector es simplemente una salida protegida que a su vez es copia directa de la señal presente en el conector MIDI-IN. En la figura (3.12) se muestra un ejemplo de interconexión mediante interfaz MIDI.

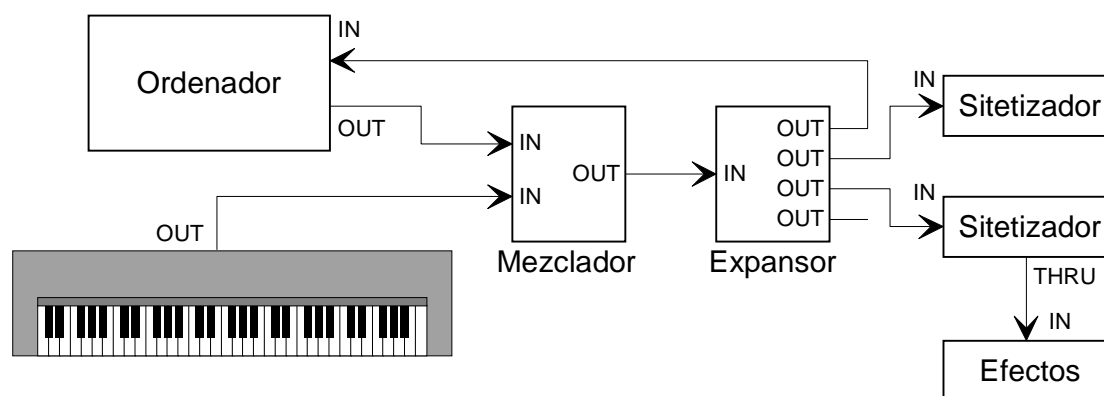


Fig. 3.12 Ejemplo de interconexión de distintos sistemas mediante interfaz MIDI

Algunos equipos incorporan fuentes de entrada múltiples con la consecuente mezcla del flujo de datos, lo cual es una operación algo compleja como se puede suponer, y exige un cuidado proceso de manera que la mezcla se realice en la secuencia adecuada. También podemos encontrar equipos con salidas múltiples, bien de tipo copia de la salida anterior, o bien tomas con la suficiente complejidad como para poder separar canales de forma autónoma. Las posibilidades de interconexión de equipos MIDI son enormes.

La configuración en cadena es la más simple que podemos formar con los equipos MIDI, y no por ello la única, o dicho de otro modo, la mejor. La configuración en anillo puede darnos excelentes resultados con equipos de la última generación, aunque también puede resultar catastrófica si son de tipo antiguo.

3.5.3 Protocolo de mensajes de MIDI

El interfaz MIDI, al contrario que el RS-232, incorpora como parte de la norma que lo define, todo el protocolo de intercambio de información entre los distintos dispositivos. Esto incluye el tipo de mensajes que intercambian los equipos y el significado de cada uno de ellos. Este protocolo es muy completo y versátil y permite grandes posibilidades en cuanto a interconexión de instrumentos musicales. Su principal inconveniente, es debido precisamente a este alto grado de normalización que lo hace esclavo del tipo de aplicaciones para las que fue pensado. Es por esto que pese a ser un interfaz económico y con múltiples posibilidades de interconexión entre equipos, como se muestra en la figura (3.12) no se emplee fuera del ámbito de la música. Sin embargo, este ámbito no se restringe únicamente a los instrumentos musicales, sino que también podemos conectar por este tipo de interfaz, generadores de efectos, mesas de mezclas digitales, equipos de grabación, bases de tiempo para sincronización de mesas de edición de audio y video o sistemas de control de iluminación, consiguiendo de esta forma efectos luminosos e incluso pirotécnicos, perfectamente sincronizados con las notas que genera un determinado intérprete en el escenario.

Aunque uno de los objetivos fundamentales de los equipos MIDI era sustituir el antiguo sistema de interfaz mediante control por voltaje, parecía lógico que sus descubridores pretendieran de él que fuese capaz de algo más que enviar el mensaje de tocar una determinada nota. Por ejemplo, mientras que los primeros sintetizadores eran casi como los órganos electrónicos de la época en que una tecla no era mucho más que un interruptor abierto o cerrado, pronto aparecieron los teclados que permitían un cierto control según la presión ejercida en ellos. Esta sensación de velocidad dio a los teclados un tacto mucho más «natural» y proporciona más «sentimiento» a la interpretación.

El protocolo MIDI está constituido por el envío de mensajes que indican al resto de equipos que realicen alguna acción, como activar o desactivar una nota concreta, cambiar el banco de sonidos, aplicar un determinado efecto, etc. Estos mensajes conforman el lenguaje a través del cual se comunican todos los dispositivos.

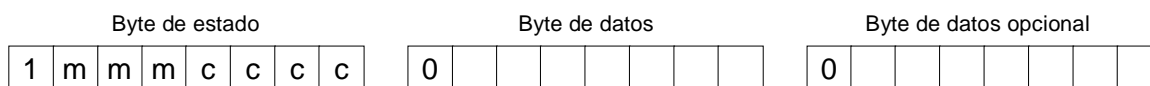


Fig. 3.13 Estructura binaria de un mensaje MIDI. Todos los mensajes comienzan por un byte de estado, que incluye 3 bits con el tipo de mensaje y 4 con el número de canal. A continuación va uno o dos bytes de datos dependiendo del mensaje que se envíe.

Los mensajes (figura 3.13) están compuestos por un byte de estado y uno o más bytes de datos. Los bytes de estado proporcionan información sobre el tipo de acción a efectuar, y seleccionan el canal en el que se realizará esa acción, mientras que los bytes de datos subsiguientes

especifican dicha acción. Por ejemplo, el byte de estado puede contener «activar una nota», y el byte de datos le indica que la nota es «un DO en la cuarta octava». En la norma MIDI los bytes de estado tienen siempre a 1 su bit más significativo (MSB), mientras que el bit más significativo de los bytes de datos siempre está a 0. La figura (3.13) muestra la estructura binaria de un mensaje MIDI.

El byte de estado incluye tres bits que especifican el tipo de mensaje, lo que permite ocho tipos distintos de mensaje. Los cuatro bits menos significativos indican el número de canal al que se dirige el mensaje. En una configuración MIDI, cada dispositivo puede ser asignado a un canal y múltiples dispositivos pueden ser asignados al mismo canal. Al hacer esto, el dispositivo responderá a los mensajes que lleven la etiqueta correspondiente a ese canal.

Byte de estado	Nº bytes de datos	Descripción	Byte de datos 1	Byte de datos 2
1000nnnn	2	Desactivación de nota	Altura	Velocidad
1001nnnn	2	Activación de nota	Altura	Velocidad
1010nnnn	2	Expresión de nota Post-pulsación	Altura	Presión
1011nnnn	2	Cambio de control	Tipo de control	Intensidad
1100nnnn	1	Cambio de programa	Programa	
1101nnnn	1	Expresión de canal	Presión	
1110nnnn	2	Cambio de tono "Pitch Bend"	MSB	LSB
1111xxxx	variable	Mensajes de sistema		

Tabla 3.3 Resumen de los distintos tipos de mensajes MIDI

Ya hemos dicho anteriormente que la mayoría de los bytes de status llevan un número de canal, y como se puede suponer, los mensajes que van precedidos por el número de canal se denominan mensajes de canal. La norma MIDI prevé también mensajes sin número de canal y destinados a causar una respuesta similar en cualquier equipo que se sitúe en el bus. Los bytes de status cuyo prefijo de mensaje lleva los cuatro bits de mayor orden a 1 son los mensajes de sistema. Es preferible que comencemos a referirnos a ellos en notación hexadecimal, por lo cual, los mensajes de sistema serán de la forma \$Fx, y donde la parte \$x será el mensaje específico (el símbolo \$ indica base hexadecimal).

En situación normal, los mensajes de canal constan de un byte de estado seguido por uno o dos bytes de datos (y cuyos MSB están a cero). El resumen de estos bytes puede verse en la tabla (3.3). El mensaje más comúnmente usado en los sistemas musicales es el que hace tocar una nota. Para ello están previstos dos bytes de status; los mensajes de canal NOTE ON y NOTE OFF. Vamos a centrarnos primero en el NOTE ON. Su codificación es \$9n, donde \$n representa el número de canal formado por los 4 bits, y requiere 2 bytes de datos. El primer byte de datos indica la nota a ser tocada. Debido a que por definición los MSB deben ser 0, hay combinaciones para 128 notas diferentes. Un piano tiene sólo 88 teclas, por lo que parecen ser suficientes, aunque algunos equipos de ultimísima generación trabajan con partituras micro-tonales, en las que se usan tonos intermedios de cada nota base del piano, por lo que en estos casos 128 combinaciones no serían suficientes. El segundo byte de datos se ocupa de especificar la velocidad, o lo que es igual, la dureza de la pulsación sobre la tecla. Si el equipo no incorpora teclado sentivo, y aunque la norma MIDI especifica que la velocidad debe ser de un valor igual a \$20, los instrumentos envían normalmente una señal cuyo valor suele centrarse sobre \$40. Es necesario que se envíen ambos bytes de datos, aunque el dispositivo no lo tenga implementado.

El mensaje de canal NOTE OFF (cuyo prefijo es \$8n) consta también de dos bytes; el número de nota y la velocidad de «tecla soltada». Pero MIDI permite que una nota sea también desactivada mediante el envío de una nueva orden de NOTE ON, cuya velocidad sea igual a 0, y la razón de ello es simplemente permitir el estado de funcionamiento continuo.

Mientras que la mayoría de los bytes de un mensaje deben ser enviados necesariamente, no ocurre lo mismo con el byte de status. Sobre todo en el caso de que el nuevo byte sea idéntico al anterior, caso en el cual la norma MIDI nos permite omitirlo. Así por ejemplo, en el caso de que desactivemos una nota mediante un mensaje NOTE ON (con V=0), no precisaremos usar los bytes de status. Si tocamos un conjunto de tres notas en el teclado, los mensajes no usarán una longitud de 18 bytes (3 triples byte de NOTE ON y 3 triples byte de NOTE OFF), sino que nos bastará con usar sólo 13 bytes. Como vemos, en caso de que todos los mensajes se sitúen en el mismo canal conseguiremos un ahorro de un 33 por 100 aproximado sobre el ancho de banda del bus. (Un «atasco» en la información enviada por MIDI puede traducirse, en la práctica, en retrasos perceptibles entre el momento de pulsar una tecla y el instante en que suena la nota correspondiente; problema éste bastante grave en los grandes sistemas, aunque hay maneras de atenuarlo).

3.6 INTERFACES PARALELO

En términos generales, una interfaz es un nexo de conexión que facilita la comunicación entre dos dispositivos.

Un gran número de unidades de disco tienen un interfaz estándar aunque algunos fabricantes, emplean interfaces no estándar en sus equipos. Afortunadamente, este tipo de interfaces propietarios y exclusivos de un fabricante son cada vez menos habituales y actualmente casi todos los dispositivos se pueden conectar a través de interfaces normalizadas, documentadas y públicas. Las interfaces las podemos dividir en dos clases; aquellas que están entre el mecanismo del dispositivo y su controlador, y aquellas que están entre la unidad básica y el controlador. El controlador puede ser una unidad separada o puede estar incluido en el dispositivo. En este último caso, la interfaz entre el dispositivo y la controladora es usualmente inaccesible y es la situación prácticamente universal hoy día. El controlador es ocasionalmente empaquetado en la unidad básica del ordenador; en este caso, el estándar 'lógico' de interfaz entre la unidad básica y el controlador se mantiene. Por lo tanto, el software puede ser usado igualmente, aunque la interfaz no exista físicamente. Aquí se entiende por controlador el circuito o dispositivo que se comunica directamente con la CPU mediante comandos y en base a esto produce unas señales de control que actúan sobre el dispositivo. Por ejemplo, un comando a un disco puede ser el leer un sector y las señales al disco son de activar un motor, mover la cabeza, comprobar posicionado, leer secuencia de patrones magnéticos y tras su decodificación enviarlos a la CPU a través del interfaz. Un comando similar podría enviarse a una impresora, sin embargo, las señales de control serían muy distintas: desplazar la cabeza de impresión, activar determinados inyectores, etc.

3.7 EL INTERFAZ ST-506/412

3.7.1 Generalidades

Entre los interfaces controladores de dispositivos uno de los primeros y más conocidos es el ST-506¹. El separador de datos está en el controlador, el cual define un rango de datos de 5 Mbits por segundo. El separador de datos es el circuito que tomando como entrada la señal proveniente de la cabeza magnética, extrae por una parte los datos y por otra la señal de reloj embebida. El

¹ Su hoja de especificaciones puede encontrarse en Peripheal INTEL vol. II

dispositivo está preparado para controlar un motor paso a paso con el fin de desplazar las cabezas pista a pista, y las pistas son seleccionadas por pulsos de motor a razón de un pulso por milisegundo. Cada uno de los pulsos mueve la cabeza a una pista, siendo el tiempo medio de acceso grande si el dispositivo tiene muchas pistas. Pueden ser seleccionadas 16 cabezas, permitiendo 4 discos. El servocontrol de la localización de la pista no se usa. Una variante de ésta es la interfaz ST-412, que contiene un buffer de pulsos. El controlador puede enviar pulsos de paso (step) más rápidamente de lo que puede desplazarse la cabeza, y el buffer los pasará, entonces, a la velocidad adecuada para el motor. De esta forma podemos usar motores paso a paso más rápidos según las necesidades. El ST-506/412 está diseñado para sistemas de bajas prestaciones y bajo costo, como los primeros PC's. Es utilizado en muchos dispositivos de 3 y 1/2" y 5 y 1/4", aunque a menudo en los PC's éstos están integrados con el controlador. Puede manejar 4 unidades de disco. Se trata de un interfaz obsoleto pero que permite ver las distintas capas de abstracción de los interfaces y la evolución hacia sistemas más modernos y mucho más completos como SCSI.

3.7.2 Cableado

Esta interfaz eléctrica está dividida en dos cables separados (etiquetados habitualmente como J1 y J2) más la alimentación (J3). J1 es un cable plano de 34 vías y J2 es también un cable plano de 20 vías. Las señales del conector J1 son de control. Todas ellas son digitales (colector abierto y niveles TTL) y funcionan en configuración simple, es decir, la información se obtiene de ellos tomando la diferencia de tensión entre la señal y una referencia común.

El cable J2 contiene señales de datos, que son diferenciales, es decir, la información se obtiene de ellos tomando la diferencia de tensión entre dos hilos de señales conjugadas. Por ejemplo, el dato leído será 1 o 0 dependiendo de la diferencia de voltaje entre *MFM_Read_Data* y *MFM_Read_Data*. Esta controladora realiza todo el trabajo y el disco lo único que hace es leer y escribir los patrones magnéticos tras acondicionar debidamente la señal. El mismo disco podría ser utilizado con codificación RLL, ya que la codificación y decodificación se realizaba en la controladora.

El conector J1 se conecta a todas las unidades en cadena, en tanto que el conector J2 se conecta por separado a cada una de ellas, como muestra la figura (3.14).

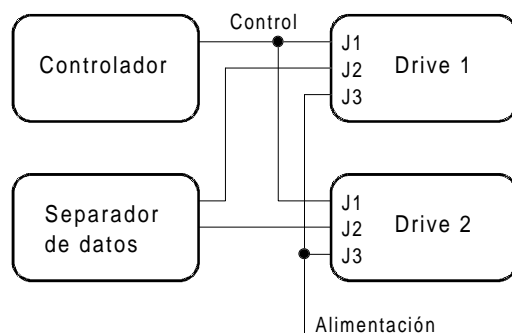


Fig. 3.14 Cableado entre el controlador, el separador de datos y los dispositivos en el ST-506/412

3.7.3 Señales y funcionalidad

CONECTOR J1

1 *Head_Select* (0:3) (O)

Estas cuatro líneas seleccionan en binario una cabeza determinada. El hecho de que sean activas a nivel bajo implica que la selección se hará con lógica negativa, es decir, la combinación 1101 en estas señales seleccionarán la cabeza número 2

- 2 $\overline{Write_Gate}$ (O)
Se activa para indicar una operación de escritura, activa la corriente de escritura en la cabeza
- 3 $\overline{Direction_In}$ (O)
Indica la dirección de la cabeza durante los pulsos de step (1 hacia afuera, 0 hacia adentro)
- 4 \overline{Step} (O)
A través de esta señal se le aplica un pulso activo a nivel bajo por cada movimiento de pista a pista que se desee mover la cabeza
- 5 $\overline{Drive_Select}$ (1:4) (O)
A través de estas cuatro señales se selecciona uno entre cuatro posibles unidades de disco que esta interfaz puede manejar. No está codificada, sino que cada disco tiene una línea, con lo cual la activación de más de una señal será considerada como activación errónea.
- 6 $\overline{Track0}$ (I)
Se activa cuando la cabeza está posicionada sobre la pista 0, que es la más exterior.
- 7 \overline{Index} (I)
Es activado por el disco cuando la cabeza está posicionada al comienzo de una pista.
- 8 \overline{Ready} (I)
Indica al controlador que en el dispositivo no existe ninguna condición de error, y por tanto, la operación en curso (lectura o escritura) puede continuar.
- 9 $\overline{Seek_Complete}$ (I)
Indica al controlador que la operación de posicionamiento sobre la pista deseada se ha completado y puede proceder a lectura o escritura.
- 10 $\overline{Write_Fault}$ (I)
Indica al controlador que la operación de escritura se ha detectado como errónea. Lo que ocurre no es que se haya escrito mal, sino que no se ha escrito debido a un fallo drástico del hardware ya que es lo único que la controladora puede detectar. Cuando un dato se ha escrito mal no se entera nadie hasta que se lee.

CONECTOR J2

- 1 $\overline{MFM_Write_Data}$ y MFM_Write_Data (O)
Esta pareja de señales constituyen el dato a escribir en el disco codificado en MFM
- 2 $\overline{MFM_Read_Data}$ y MFM_Read_Data (O)
A través de estas dos señales el disco envía a la controladora la secuencia de bits leídos

3 *Drive_Selected* (I)

Con esta señal, el dispositivo indica al controlador que ha reconocido la selección. Es lógico que esta señal no forma parte del conector J1, puesto que notifica una situación particular de cada cable. Si al controlador le llegara una señal de *Drive_Selected* por el cable J1 (que está conectado a todos los dispositivos) no sabría quién le está contestando.

3.7.4 Ejemplo de implementación: La tarjeta controladora WD1003-WAH

Aquí se verá un ejemplo de una tarjeta controladora para la interfaz ST-506/412. El diagrama de bloques de esta tarjeta puede verse en la figura (3.15)

La solución que se comenta, como la mayoría de ellas, se basa en el uso de un conjunto de circuitos integrados, que consiste en dividir la función a realizar en cometidos más simples que puedan ser abordados por integrados de no muy alta complejidad. Estos chips están pensados para conectarse entre sí casi pin a pin. El resultado es que se tiene una tarjeta con 4 ó 5 chips LSI o VLSI (que realizan entre todos la función compleja) y varios chips SSI (puertas lógicas y 'drivers' excitadores de línea). Actualmente estas controladoras tienden a realizarse a medida con lo que se englobaría todo en un solo chip. La tarjeta en cuestión es la WD1003-WAH y controla hasta dos discos winchester de hasta 16 cabezas y 2048 cilindros cada uno. La interfaz con la unidad básica que presenta esta tarjeta, es el slot de expansión de los PC's.

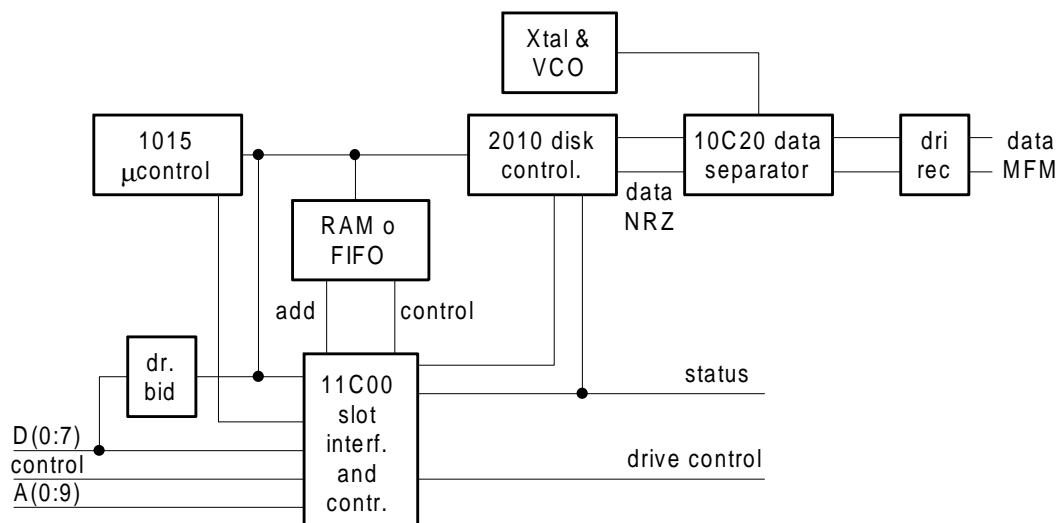


Fig. 3.15 Diagrama de bloques de una tarjeta controladora para el ST-506/412

Está basada en el conjunto de integrados formado por los siguientes circuitos:

1 WD2010A-05

Es el controlador de discos winchester. Suministra la mayoría de las señales de la interfaz ST-506. La única función que deja para otro chip es la de separador de datos (realizada por el WD10C20). Por el lado de la unidad básica se conecta con una memoria tampón (una FIFO o una RAM y un contador) y un controlador de esta memoria. Además suministra el control de interrupciones y transferencias por DMA.

2 WD11C00C-22

Es el controlador de la interfaz con la unidad básica, en este caso el PC. Además de esta interfaz física, ejecuta algunas otras funciones como el direccionamiento del buffer para las transferencias en escritura con la unidad básica, separa en bytes la palabra (word) que le envía la unidad básica, realiza él la selección de cabeza, controla el led de actividad del disco, etc.

3 WD1015-27

Es un microcontrolador de 8 bits que controla y coordina el buen funcionamiento de los dos LSI anteriores. Recibe y envía información de comandos y de estado a través del bus interno (que está multiplexado) de la tarjeta. El 'firmware' de control reside en una ROM interna de 2K que posee el propio micro.

4 WD10C20

Es el separador de datos. En lectura, realiza la sincronización de los datos del disco, suministra al controlador (WD2010A-05) el reloj con el que debe muestrear los datos para decodificar la secuencia MFM. En escritura, realiza la precompensación. Hay que suministrarle por el exterior algunos elementos analógicos para el VCO (Oscilador controlado por voltaje) interno.

Esta tarjeta presenta a la unidad básica 7 registros de lectura/escritura internos al WD2010A-05 y otros 3 más externos. Los registros internos al controlador se llaman registros de tarea. En ellos se escribirán los comandos y los parámetros de las operaciones a realizar. Como ejemplo del tipo de información que manejan estos registros, se enumeran en la tabla (3.4) los registros de tarea.

Dirección	Lectura	Escritura
1F1 h	Registro de errores	Precompensación
1F2 h	Cálculo del sector	Cálculo del sector
1F3 h	Cilindro (byte bajo)	Cilindro (byte bajo)
1F4 h	Cilindro (byte alto)	Cilindro (byte alto)
1F5 h	Dispositivo/Cabeza	Dispositivo/Cabeza
1F6 h	Estado	Comando

Tabla 3.4 Registros del controlador WD2010A-05

El controlador soporta 8 comandos posibles, que se escribirán para su ejecución en el registro 1F6h. Los comandos son los siguientes:

1 *Restore*

Mover la cabeza hasta la pista 0

2 *Seek*

Mover la cabeza un número de pasos determinado

3 *Read Sector*

Leer un número de sectores de 1 a 256 a partir de uno dado (el número es el Sector Count)

4 *Write Sector*

Análogo al anterior pero de escritura.

5 *Read Verify*

Leer un sector para la comprobación del CRC sin pasarlo a la unidad básica

6 *Format Pista*

Formatear una pista

7 *Diagnose*

Ejecuta una rutina de autotest de la tarjeta y comunica el resultado a la unidad básica

8 *Set Parameter*

Fijar parámetros, como el número de sectores por pista, el número de cabezas, etc.

La ejecución típica de comandos se lleva a cabo con la siguiente secuencia de operaciones:

- En reposo, el controlador tiene las señales de control desactivadas y su registro de estado indica la situación de Ready
- La unidad básica escribe los parámetros de la operación a realizar en los registros de tarea. Luego escribe el comando. Si la operación es de escritura, debe escribir el sector a escribir en la memoria tampón. En este caso el WD11C00C (el controlador de la interfaz con la unidad básica) sabe que el comando requiere que se rellene el buffer y controla esta transferencia por DMA
- Cuando termina esta transferencia (si alguna vez empezó), el microprocesador (WD1015-27) notifica al controlador que puede empezar la ejecución del comando, pues todo está listo.
- A la terminación del comando el controlador lanza una interrupción al microcontrolador, éste examina el registro de estado y si decide que la información es coherente con el comando que se ejecutó, entonces lanza a la unidad básica la interrupción y se retorna al estado de reposo.

Aunque este esquema de funcionamiento pueda parecer farragoso, responde al criterio de división de tareas y responsabilidades. Este modo de operar tiene la ventaja de que se puede mejorar el sistema mejorando cualquiera de sus componentes. Por ejemplo, se puede cambiar el circuito codificador MFM por otro RLL manteniendo el mismo disco con lo la capacidad se verá aumentada en un 50% como ya se comentó en el capítulo anterior.

3.8 INTERFAZ ESDI

Posteriormente las investigaciones han ido encaminadas a dispositivos que incorporen el separador de datos, porque esto da al diseñador más posibilidades para incrementar la capacidad y el rendimiento. El más popular de los dispositivos para control de pequeños discos con separador de datos es el "Extended Small Device Interface" o ESDI. Esto no restringe el rango de datos, y la mayoría de los controladores manejan cualquier rango sobre 10 Mbits por segundo (algunos 20 Mbits/seg). La selección de la pista se realiza transmitiendo por las direcciones a razón de un pulso por pista. Hay previsión para que el controlador interroge al dispositivo, el cual informará sobre los parámetros necesitados por el controlador tales como número de cabezas, cilindros, bytes por pista, etc.

3.9 BUS SCSI

3.9.1 Generalidades

La interfaz entre la unidad básica y el controlador es a menudo diseñada para seguir las especificaciones de la unidad básica (por ejemplo, el IBM PC). Sin embargo, existe actualmente una tendencia hacia el uso de interfaces independientes de la unidad básica estándar, de los cuales la más popular es la 'Small Computer System Interface' o SCSI, que es un desarrollo del SASI producido por la firma 'Shugart Associates Standard Interface'. SCSI está ahora reconocida como estándar ANSI bajo el nombre ANSI X3.131-1986. Soporta muchos tipos de periféricos, no sólo discos, y con simples adaptadores puede conectarse a la mayoría de las unidades básicas. Algunas unidades básicas de computadora recientes soportan SCSI y no necesitan ni siquiera adaptador.

SCSI puede ser usada de dos maneras: como interfaz de la unidad básica construida en una interfaz inteligente, y como interfaz de la unidad básica con un controlador separado, que conecta uno o más dispositivos. La primera forma es normalmente la más barata cuando se utiliza un único dispositivo. La segunda forma se utiliza cuando hay un grupo de dispositivos (o una mezcla de dispositivos periféricos). La especificación SCSI tiene muchas características opcionales. Una de ellas permite al controlador dirigirse al dispositivo para iniciar un movimiento de la cabeza, y entonces desconectar el dispositivo del controlador y conectarlo a otro dispositivo cuya cabeza está ya en la pista correcta: puede transferir datos durante un tiempo comparativamente largo antes de que el primer dispositivo esté preparado para realizar eso. Cuando hay más de un programa, o más de una unidad básica, usando un grupo de discos, este rasgo puede aumentar el rendimiento del sistema. Sin embargo, el gran número de opciones y alternativas permitidas por el estándar SCSI puede dar problemas de compatibilidad puesto que algunos fabricantes no implementa un desarrollo completo de la norma.

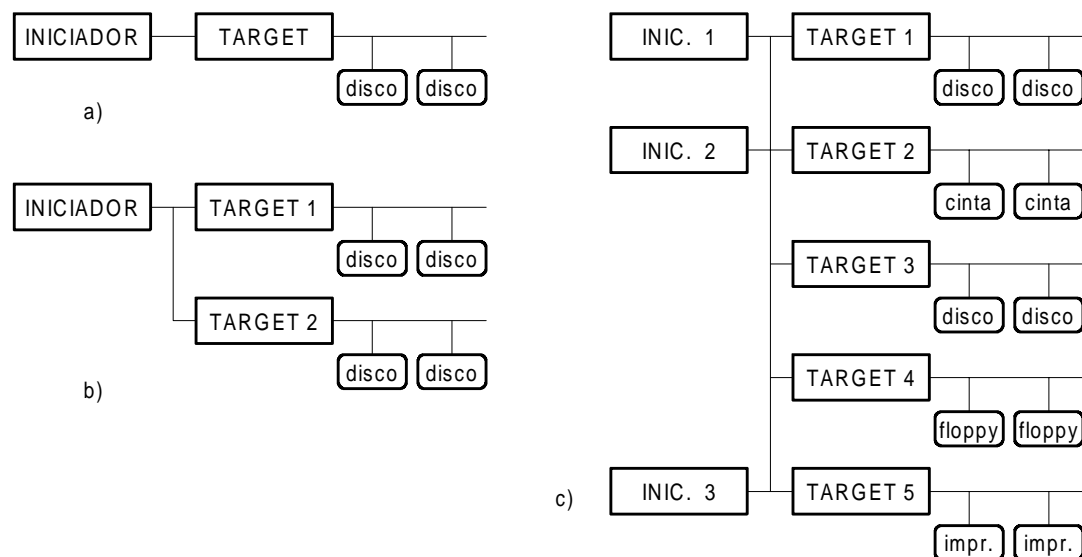


Fig. 3.16 Algunas configuraciones permitidas por la norma SCSI. a) Un único iniciador y un único target, b) Un único iniciador y varios target, c) Varios iniciadores y varios target. En todos los casos se han incluido varios dispositivos lógicos en cada periférico.

Según especifica la norma, existen dos tipos de dispositivos que se pueden conectar al bus SCSI, el iniciador (generalmente un procesador principal) y el target (generalmente un periférico). Al bus se pueden conectar un máximo de 8 dispositivos (iniciadores o targets), de los cuales al menos uno debe ser un iniciador y al menos uno debe ser un target. En la figura (3.16) se muestran algunas configuraciones posibles. Un determinado periférico SCSI puede contener internamente varios dispositivos, como por ejemplo varios discos, o varios lectores de CD-ROM y ser considerados como un único periférico por el sistema central. Estos elementos, reciben el nombre

de dispositivos lógicos y se les asigna un segundo identificador o LUN (Logic Unit Number). Dentro de un periférico SCSI que tiene varios dispositivos lógicos, éstos deben ser iguales y aceptar los mismos comandos. Salvo en los grandes sistemas, un periférico SCSI tiene normalmente un único dispositivo al que se le asigna el LUN cero.

Toda comunicación (síncrona o asíncrona) a través del bus se realiza siguiendo una secuencia determinada de eventos. Estos eventos están agrupados por la funcionalidad en lo que se denominan fases. El bus en cada instante sólo se encontrará en una fase concreta. La fase está determinada por el estado de las señales de control del bus.

A continuación se dará una relación de las señales y la funcionalidad de la interfaz SCSI. Luego se describen someramente cada una de las fases, sin hacer referencia a los tiempos que especifica la norma², considerando en primer lugar la transferencia asíncrona de datos y la transferencia síncrona de datos, así como la descripción de las fases de transferencia de información. Finalmente, se describen las condiciones especiales del bus.

3.9.2 Señales y funcionalidad

Esta interfaz tiene un total de 18 señales. Nueve son usadas para el control y las nueve restantes son para datos. El bus de datos es por tanto paralelo de 8 bits más uno de paridad.

Existen dos modos de implementación eléctrica, una simple (con referencia a una tierra común) o 'single-ended' (SE) y otra diferencial o 'differential-ended' (DE) donde cada señal es transmitida por un par de hilos conjugados. En la configuración simple, el nivel lógico se obtiene de la diferencia de potencial entre el hilo de señal y la tierra común. En la configuración diferencial, se obtiene de la diferencia de potencial entre cada par de hilos conjugados. El modo simple es más cómodo y barato de implementar pero está limitado a un cable de 6 metros (que es suficiente si se implementa como bus local). El cable diferencial puede llegar hasta 25 metros. Estas dos implementaciones son incompatibles entre sí, y por lo tanto, en un sistema todos los dispositivos deberán ser del mismo tipo, no pudiéndose mezclar dispositivos 'single-ended' con dispositivos 'differential-ended'. No obstante, en un mismo sistema pueden convivir más de un bus con lo que un sistema puede trabajar simultáneamente con dispositivos SE y DE siempre que se conecten a buses distintos. En los sistemas pequeños o medianos, habitualmente sólo se emplea conexión SE.

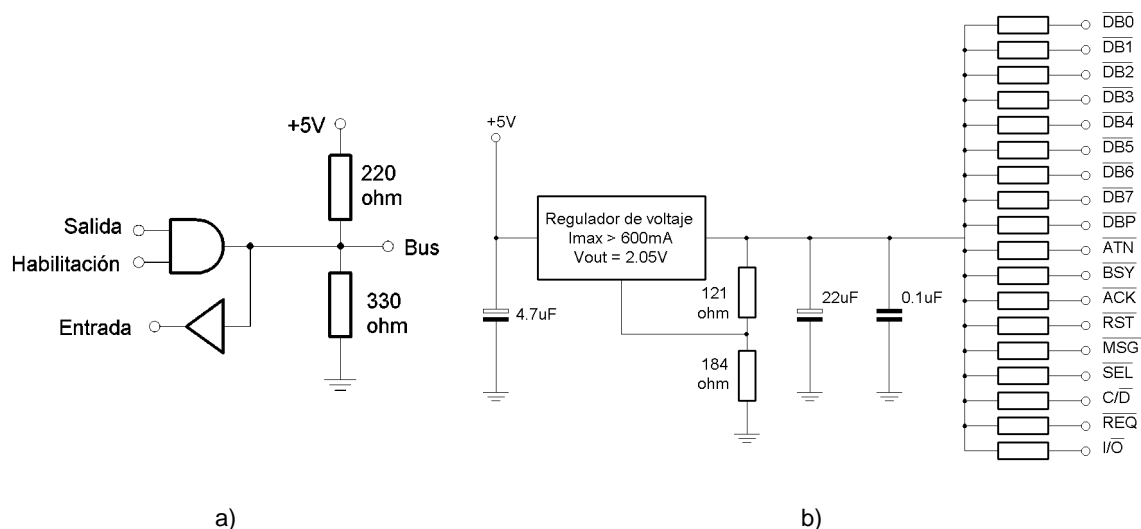


Fig. 3.17 Circuitos terminadores del bus en modo simple. a) en SCSI-1 b) alternativa en SCSI-2

² Aquellos que estén interesados en los tiempos pueden dirigirse a las normas ANSI X3.131-1986 (SCSI-1) y ANSI X3.131-1994 (SCSI-2)

Tanto una como otra implementación deben incorporar un terminador en los extremos del bus. Estos terminadores son necesarios ya que al trabajar a frecuencias elevadas, hay que tener en cuenta las posibles reflexiones de señal que se pueden producir en los extremos de los cables causadas por una desadaptación de impedancias. Hay que tener en cuenta que cuando por una línea eléctrica se propagan señales de alta frecuencia, no pueden considerarse simplemente señales, sino que deben considerarse como ondas electromagnéticas que se propagan a través de una línea de transmisión. Al comportarse como ondas, pueden producirse reflexiones en los extremos de un cable de la misma forma que se reflejan las ondas de agua en las paredes de un estanque. Un estudio detallado de este fenómeno requeriría entrar en la teoría de líneas de transmisión y queda por tanto fuera del alcance de este curso. Basta tener en cuenta que si se produjesen reflexiones en la señal transmitida a través del bus, podrían producirse interferencias entre la señal emitida y la reflejada desde el extremo lo que provocaría un mal funcionamiento del bus. Los terminadores tienen por tanto la función de simular la continuación del bus como si de una línea infinita se tratase. En las figuras (3.17) y (3.18) se muestran los circuitos terminales que deben colocarse en los extremos del bus para una y otra configuraciones.

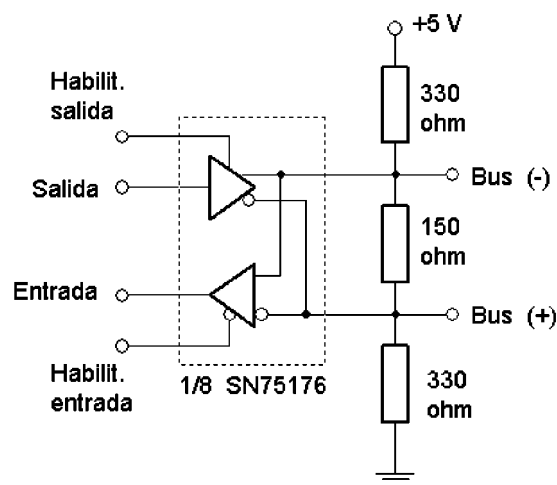


Fig. 3.18 Circuito terminador para SCSI en modo diferencial.

Las señales del bus SCSI son las siguientes:

- 1 \overline{BSY}
La señal *busy* indica que el bus está siendo usado. Cuando \overline{BSY} está activada, nadie puede acceder al bus excepto los dos dispositivos interlocutores.
- 2 \overline{SEL}
La señal *select* la emplea un iniciador para seleccionar un target o bien un target para seleccionar un iniciador.
- 3 C/\overline{D}
La señal *control/data* la maneja el target para indicar si por el bus de datos circula información de control o de datos.
- 4 I/\overline{O}
La señal *input/output* la maneja el target para indicar la dirección de los datos en el bus (cuando es verdadera indica entrada al iniciador). Esta señal es usada también para distinguir entre las fases de selección y reselección.
- 5 \overline{MSG}
La señal *message* es la señal con la que el target indica que la información en curso en el bus es un mensaje.

- 6 \overline{REQ}
La señal de *Request* (solicitud o petición) es la que el target emplea para indicar una petición en el protocolo de transferencia de datos por *REQ/ACK*
- 7 \overline{ACK}
La señal *Acknowledge* (reconocimiento) es la que maneja el iniciador para indicar un reconocimiento en el protocolo de transferencia de datos *REQ/ACK*
- 8 \overline{ATN}
La señal *Attention* es la señal con la cual el iniciador indica la condición de atención.
- 9 \overline{RST}
La señal *Reset* indica la condición de reset. Puede ser activada por cualquier dispositivo en cualquier momento y coloca al bus en un estado inicial.
- 10 \overline{DB} (0:7), *DBP*
Las señales *Data Bus* son ocho señales de datos y una señal de paridad y conforman el bus de datos. *DB7* es el bit de más peso (más significativo). La paridad *DBP* de los datos es impar. El empleo de la paridad es una opción del sistema.

El cable de conexión es una cinta plana de 50 vías, que está conectada en cadena a todos los dispositivos.

3.9.3 Fases del bus SCSI

FASE DE BUS LIBRE

Se emplea para indicar que ningún dispositivo se encuentra utilizando el bus y que se encuentra disponible para los demás. Los equipos conectados no detectan la fase de bus libre hasta que \overline{SEL} y \overline{BSY} permanecen falsos durante un cierto tiempo (>400ns)

A esta fase se llega cuando se conecta la alimentación o después de un RESET del bus. También se llega a esta fase durante una operación normal del bus si se envía alguno de los mensajes: COMMAND COMPLETE, DISCONNECT, ABORT, BUS DEVICE RESET, RELEASE RECOVERY, ABORT TAG y CLEAR QUEUE.

FASE DE ARBITRAJE

La fase de arbitraje permite a un dispositivo (iniciador o target) ganar el control del bus para que pueda comunicarse con otro (target o iniciador).

La implementación de la fase de arbitraje es una opción del sistema. Aquellos que no implementan esta opción tienen un solo iniciador y un solo target, que siempre estarán lógicamente conectados al bus. Para ganar el bus, cada dispositivo tiene asignado un identificador que coincidirá con uno de los bits del bus de datos, de ahí que sólo se puedan conectar 8 dispositivos al bus. En el arbitraje, si dos o más dispositivos colisionan para obtener el control, siempre lo obtendrá el que tenga el identificador de más peso.

El procedimiento que sigue un dispositivo para obtener el control del bus SCSI es como sigue:

- El dispositivo esperará primero a que ocurra un fase de bus libre. La fase de bus libre se detecta cuando \overline{BSY} y \overline{SEL} son ambas simultánea y continuamente falsas durante cierto tiempo (>400 ns).
- El dispositivo esperará otro retraso después de la detección de bus libre y antes de introducir ninguna señal.
- A continuación, el dispositivo ya puede arbitrar introduciendo la señal \overline{BSY} y las correspondientes a su propio identificador.

- Después de esperar un cierto tiempo, el dispositivo examinará el bus de datos. Si un bit ID SCSI de mayor prioridad es verdadero, entonces el dispositivo ha perdido el arbitraje y tiene que desactivar sus señales y volver a esperar a que ocurra una fase de bus libre. Si no hay un bit ID más prioritario, entonces el dispositivo ha ganado el arbitraje y activará \overline{SEL} . Los demás dispositivos que participaron en el arbitraje han perdido y desactivan sus señales \overline{BSY} y su ID y volverán a esperar a que ocurra una fase de bus libre. El bit de paridad no es válido durante el arbitraje.

FASE DE SELECCIÓN

La fase de selección permite a un iniciador seleccionar un target con el propósito de iniciar alguna función. Se llega a esta fase cuando la fase de arbitraje ha sido ganada por un iniciador. Si es ganada por un target, se entra en la fase de RESELECCIÓN, existiendo por tanto una simetría entre ambas fases.

El dispositivo del bus SCSI que gana el arbitraje tiene \overline{BSY} y \overline{SEL} validadas; a continuación pondrá el bus de datos a un valor que será la OR de su bit y el bit ID del seleccionado. De esta forma el dispositivo seleccionado puede saber quien lo ha hecho. El dispositivo esperará un cierto tiempo y desactivará \overline{BSY} .

El seleccionado se enterará de la selección cuando \overline{SEL} y su bit ID estén activos, y \overline{BSY} inactiva. El seleccionado podrá examinar el bus de datos para determinar el ID del ganador del bus (a menos que se emplee la opción de no arbitraje, en cuyo caso se sabe quién le habla) y activará \overline{BSY} para indicar al seleccionador (dispositivo que ganó el arbitraje) que se ha dado cuenta que ha sido seleccionado y por quien. Si hay más de dos bits en el bus de datos no se responderá a la selección.

Después de que el seleccionador detecta que \overline{BSY} es verdadera, liberará \overline{SEL} y podrá continuar el proceso. En la figura (3.19) se esquematiza el secuenciamiento de una fase de selección.

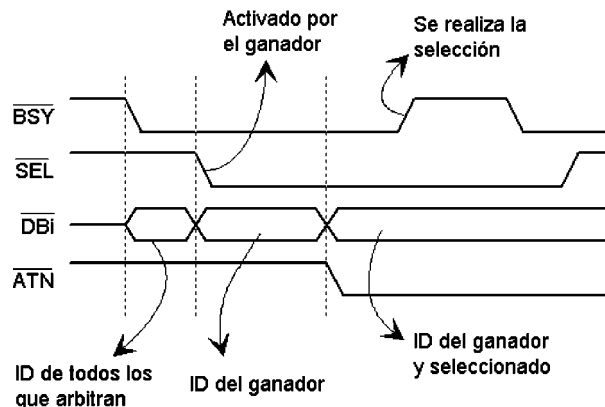


Fig. 3.19 Cronograma de una fase de selección tras la fase de arbitraje

FASE DE RESELECCIÓN

La reelección, al igual que la de selección, sólo puede ser usada en sistemas que tienen la fase de arbitraje implementada. Sólo se diferencia de la fase de selección en que el ganador ha sido un target y va a seleccionar a un iniciador.

Para avisar al iniciador que le selecciona un target, éste activa I/\overline{O} en esta fase, y permanecerá activa hasta el final de la misma. Salvo esta señal, el resto es idéntico a la fase de selección.

3.9.4 Fases de transferencia de información

Los conceptos de entrada/salida están referidos aquí siempre al iniciador. Es decir cuando se habla de entrada (salida) de datos se entiende que los datos son enviados por el target (iniciador) y recibidos por el iniciador (target).

Las señales C/\overline{D} , I/\overline{O} y \overline{MSG} son usadas para distinguir entre las distintas fases de transferencia de información. El target establece estas tres señales y por tanto controla todos los cambios de una fase a otra. El iniciador se ve obligado a responder a ellas. Lo único que le está permitido es provocar una reinicialización o una condición de atención, que se verá más adelante en el apartado dedicado a las condiciones especiales del bus.

Las fases de transferencia de información emplean un protocolo *REQ/ACK* para controlar la comunicación. En cada *REQ/ACK* se transfiere un byte de información. Durante las fases de transferencia de información \overline{BSY} quedará activa para que nadie interfiera la comunicación. En la tabla (3.5) se resume la codificación de las fases de transferencia de información.

\overline{MSG}	C/\overline{D}	I/\overline{O}	Nombre de fase	Dirección
1	1	1	Salida de Datos	Iniciador → Target
1	1	0	Entrada de Datos	Target → Iniciador
1	0	1	Comando	Iniciador → Target
1	0	0	Estado	Target → Iniciador
0	1	1	*	Reservado
0	1	0	*	Reservado
0	0	1	Salida de Mensaje	Iniciador → Target
0	0	0	Entrada de Mensaje	Target → Iniciador

Tabla 3.5 Tabla de fases de transferencia de información.

FASE DE COMANDOS

La fase de comandos permite al target pedir un comando al iniciador. El target activará la señal C/\overline{D} y negará I/\overline{O} y \overline{MSG} durante los intercambios *REQ/ACK* de esta fase.

FASE DE DATOS

Es un término que engloba la fase de entrada de datos y la fase de salida de datos.

- La fase de entrada de datos permite al target pedir que el iniciador acepte datos desde el target. El target activará la señal I/\overline{O} y negará las señales C/\overline{D} y \overline{MSG} durante esta fase.
- La fase de salida de datos permite al target pedir que los datos sean enviados desde el iniciador al target. El target negará las señales C/\overline{D} , I/\overline{O} y \overline{MSG} durante el protocolo *REQ/ACK* de esta fase.

FASE DE ESTADO

Permite al target enviar información de estado (normalmente de éxito o fracaso de la ejecución de un comando) al iniciador. El target activará C/\overline{D} e I/\overline{O} y negará la señal \overline{MSG} durante esta fase. La información de estado siempre es de 1 byte. Al contrario de un mensaje que puede ser enviado en cualquier momento durante la fase de comando, el estado solo se envía cuando el comando se ha completado, se ha interrumpido o ha sido rechazado.

FASE DE MENSAJE

La fase de mensaje también engloba la de entrada de mensaje y la de salida de mensaje

- La fase de entrada de mensaje permite al target pedir que el iniciador acepte un mensaje, es decir, información de control no ejecutable ni que es consecuencia directa de la ejecución de un comando. El target activará C/\overline{D} , I/\overline{O} y \overline{MSG} .
- La fase de salida de mensajes permite al target pedir que el mensaje sea enviado desde el iniciador al target. El target activará C/\overline{D} y \overline{MSG} y negará I/\overline{O} .

El iniciador puede solicitar esta fase provocando la condición de atención.

Como resumen de todas las fases, en la tabla (3.6) se muestra el origen de las señales en todas ellas. En la figura (3.20) se muestra el diagrama de transición entre las distintas fases del bus. Este esquema puede interpretarse como un diagrama de transición de estados y diseñar su circuito secuencial equivalente.

Nombre de fase	\overline{BSY}	\overline{SEL}	$C/\overline{D}, I/\overline{O},$ $\overline{MSG}, \overline{REQ}$	$\overline{ACK}, \overline{ATN}$	$\overline{DB0} - \overline{DB7},$ \overline{DBP}
Bus libre	Nadie	Nadie	Nadie	Nadie	Nadie
Arbitraje	Todos	Ganador	Nadie	Nadie	ID
Selección	I y T	Ganador	Nadie	Iniciador	Iniciador
Reselección	I y T	Target	Target	Iniciador	Target
Comando	Target	Nadie	Target	Iniciador	Iniciador
Entrada de datos	Target	Nadie	Target	Iniciador	Target
Salida de Datos	Target	Nadie	Target	Iniciador	Iniciador
Estado	Target	Nadie	Target	Iniciador	Target
Entrada de mensaje	Target	Nadie	Target	Iniciador	Iniciador
Salida de mensaje	Target	Nadie	Target	Iniciador	Target

Tabla 3.6 Origen de las señales en cada una de las fases.

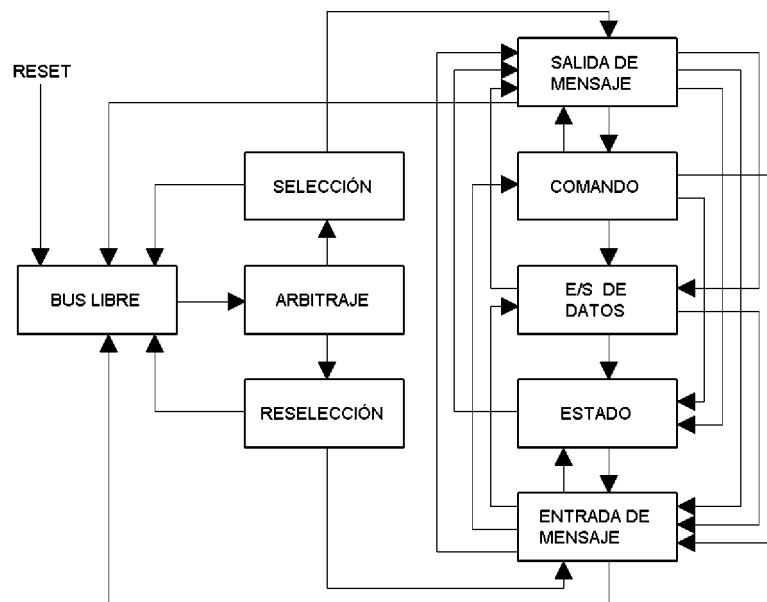


Fig. 3.20 Transición de fases del bus SCSI

3.9.5 Variantes síncrona y ancha

Existe la posibilidad de un modo de transferencia en la cual el protocolo *REQ/ACK* se modifica para que sea más rápido. Este modo es opcional y se le conoce como SCSI rápido (Fast SCSI) o modo síncrono.

Para poder utilizarla, ambos dialogantes se deben haber puesto previamente de acuerdo en una serie de cosas usando mensajes. Este modo consiste básicamente en que no se espera el *ACK* del iniciador para continuar con el siguiente dato. Como existe acuerdo de anchura mínima y separación mínima entre pulsos y número máximo de \overline{REQ} sin esperar *ACK*, se envía un tren de pulsos y bytes sin esperar contestación y se van contando las contestaciones recibidas. Si al final coinciden se podrá continuar. Con este modo se consigue subir la velocidad de transferencia de 1.5 Mbytes/s hasta 4 Mbytes/s.

Existe una segunda variante del interfaz SCSI conocida como Wide-SCSI. Este interfaz es formalmente idéntico al SCSI normal, pero en lugar de trabajar con un bus de 8 bits de datos y uno de paridad, trabaja con 16 bits de datos y 2 de paridad o 32 bits de datos y 4 de paridad (1 bit de paridad para cada byte de la doble palabra de 16 o 32 bits).

Por compatibilidad con el SCSI normal, un segundo cable es añadido para llevar todas las señales adicionales. A este segundo cable se le llama cable B para distinguirlo del cable normal que en los sistemas Wide SCSI se denomina cable A. Este segundo cable (o cable B), consta de 68 hilos con sus correspondientes conectores. Actualmente, el comité de normalización encargado del bus SCSI propone sustituir estos dos cables por uno para transferencias de 8 o 16 bits y otro alternativo para transferencias a 32 bits con lo que el segundo cable queda eliminado en las últimas versiones de la norma.

En un mismo sistema se pueden mezclar dispositivos SCSI normales con dispositivos Wide-SCSI. Como las señales adicionales van por otro cable independiente, puede suceder que su longitud difiera de las del bus SCSI normal, por lo que en este cable adicional se han incluido dos señales nuevas, aparte de las ya mencionadas de datos y paridad: \overline{REQB} y \overline{ACKB} que tienen el mismo significado que \overline{REQ} y \overline{ACK} en el cable A, pero referidas a las señales que van por el cable B. Las señales \overline{REQ} y \overline{REQB} solicitan ('request') algún tipo de atención y \overline{ACK} y \overline{ACKB} envían el reconocimiento de llegada del dato ('acknowledgement'). Con la eliminación del segundo cable, esto ya no es necesario.

3.9.6 Condiciones especiales del bus

El bus SCSI tiene dos condiciones asíncronas y que por lo tanto quedan fuera del esquema de fases síncronas descrito anteriormente y esquematizado en la figura (3.20): la condición de ATENCIÓN y la condición de RESET. Estas condiciones provocan que el dispositivo lleve a cabo algunas acciones peculiares y se altere la secuencia de fases.

CONDICIÓN DE ATENCIÓN

Permite a un iniciador informar al target de que tiene un mensaje preparado. El target debe responder con la fase de salida de mensaje.

El iniciador crea una condición de atención activando \overline{ATN} en cualquier momento, excepto en las fases de arbitraje y de bus libre. El iniciador niega \overline{ATN} durante el último intercambio *REQ/ACK*.

CONDICIÓN DE RESET

Se utiliza para realizar un reset. Prevalece sobre todas las demás fases y condiciones. Cualquier dispositivo puede provocar un reset simplemente activando la señal \overline{RST} , el estado de todas las señales del bus SCSI distintas de \overline{RST} mientras esta señal está activa no está definido.

3.10 LOS INTERFACES CENTRONICS E IEEE-1284

3.10.1 Introducción y necesidad de la norma

Uno de los interfaces más extendidos en todos los sistemas computadores, e implementados en gran número de dispositivos periféricos es el interfaz de tipo paralelo Centronics. En algunos casos se le conoce simplemente como puerto paralelo, aunque existen otros interfaces de tipo paralelo distintos y ampliamente difundidos, como SCSI o GPIB. El tipo de dispositivos que han incluido, y siguen incluyendo este tipo de interfaz son los dispositivos de salida, que podríamos catalogar como lentos, fundamentalmente impresoras y plotters. Esto es debido a que la sencillez del hardware de este interfaz, no permitía grandes posibilidades de gestión de la comunicación, y por lo tanto, todo el trabajo debía recaer sobre un programa de control. Esto hace que el protocolo sea tremendamente lento, si tenemos en cuenta que se trata de un interfaz implementado con un bus paralelo de 8 bits. Actualmente el protocolo se ha mejorado con una mayor velocidad lo que permite ampliar el tipo de periféricos conectables mediante este tipo de interfaz.

Las características fundamentales de este interfaz, son que se trata de un interfaz con bus de datos paralelo (8 líneas) un bus de control (4 líneas) y 5 líneas de estado, cuyos significados están intimamente relacionados con las impresoras. Es fundamentalmente unidireccional, y punto a punto, es decir que solo puede comunicar un ordenador con un periférico, ya que no implementa la posibilidad de conectar más dispositivos simultáneamente, por lo que no se puede emplear el término de 'bus' para referirnos a él como sí sucede con SCSI o GPIB.

Cuando IBM introdujo en el mercado el PC, a mediados de 1981, incluyó el puerto paralelo como una alternativa rápida al interfaz serie, que era el más habitual para trabajar con impresoras, y terminales de datos. Este aumento de velocidad se debía al hecho de que Centronics es un interfaz paralelo en lugar de serie, lo que permitía enviar un byte (8 bits) completo de datos cada vez. Por contra tiene el inconveniente de no permitir distancias elevadas entre el ordenador y el dispositivo periférico. Las distancias máximas, así como otras características no fueron recogidas en ninguna norma estándar, por lo que los fabricantes lo han implementado con ciertas variaciones, que si bien no son muy importantes, si son muy numerosas. La influencia de estas numerosas variantes ha sido relativamente pequeña debido fundamentalmente al bajo nivel de requisitos que se le ha exigido a este tipo de interfaz, como lo demuestra el hecho de que se haya empleado casi exclusivamente para enviar datos a impresoras u otros dispositivos de salida con tiempos de respuesta elevados, y que habitualmente están próximos al sistema computador.

No obstante, cuando comenzó a popularizarse el uso de los ordenadores portátiles surgió la necesidad de mejorar este interfaz para hacerlo útil en el uso de otros tipos de dispositivos de mayor velocidad y nivel de requerimientos. Esto fue debido a que, mientras que en un sistema central, no es muy problemático añadir un interfaz específico, en un sistema portátil, si que lo es. Por este motivo se empezaron a diseñar dispositivos periféricos que empleasen el puerto paralelo Centronics, ya que está presente en casi todos los sistemas, para añadir cualquier otro tipo de periféricos, como por ejemplo discos duros, CD-ROM, conexiones a red local, etc..

Esto provocó que el interfaz entrara en una fase de mejoras, donde cada fabricante introducía las suyas propias. Estas mejoras se encaminaban a dos objetivos fundamentales: aumentar la velocidad de transferencia, y conseguir un interfaz verdaderamente bidireccional ya que estas eran las dos grandes limitaciones del interfaz. Estas mejoras individualistas no tardaron en traer como consecuencia problemas de compatibilidad, lo que provocó que varias empresas

como Lexmark, IBM, Texas Instrument y otras, fuertemente introducidas en el campo de la microinformática y por extensión, en la informática portátil, hicieran frente común y crearan la Network Printing Alliance (NPA). Esta asociación de fabricantes definió una serie de parámetros para permitir una completa comunicación entre impresoras y sistemas computadores. Un requisito que consideró esencial fue el de dotar al interfaz de un modo verdaderamente bidireccional, lo que permitiría ampliar considerablemente el tipo de dispositivos que se podrían conectar. Una restricción era que la nueva interfaz debía ser compatible con la interfaz Centronics existente, debido al gran número de sistemas que lo estaban empleando. Prácticamente todos los ordenadores disponían de uno ya que su sencillez hacía que el coste del mismo fuera muy bajo. Además la gran mayoría de impresoras en funcionamiento no disponían de ningún otro tipo de conexión por lo que un cambio significativo las dejaría 'desconectadas'.

Esta asociación propuso al IEEE (Institute of Electrical and Electronic Engineering) la creación de un comité para desarrollar un nuevo estándar para el 'puerto paralelo' que lo hiciese de alta velocidad y bidireccional, llegando al menos a 1Mbyte por segundo de velocidad de transferencia en ambas direcciones. Este comité fue el IEEE-1284, por lo que el nuevo estándar para el puerto paralelo se conoce como IEEE-1284.

Este estándar define cinco modos de funcionamiento, el más simple de los cuales corresponde con el interfaz Centronics convencional y recibe el nombre de modo compatible (Compatibility mode). El resto son el modo Nibble, de Byte y los modos de altas prestaciones EPP (Enhanced Parallel Port) y ECP (Extended Capability Port). Estos dos últimos, elevan considerablemente la velocidad de transferencia, ya que gestionan el protocolo por hardware. En el interfaz Centronics convencional, es el propio software el que se encarga de gestionar completamente la comunicación, activando y desactivando por programa todas las señales de control. A continuación se dará una breve descripción de los distintos modos. Hay que señalar que la norma designa con distintos nombres a las señales en cada uno de los modos ya que su significado varía de un modo a otro.

Grupo	Señal SPP	E/S	Descripción
Control	nSTROBE	S	Activa en baja. Indica que hay un dato válido en las líneas de datos.
	nAUTOFEED	S	Activa en baja. Indica a la impresora que añada un avance de línea de forma automática por cada retorno de carro
	nSELECTIN	S	Activa en baja. Usada para indicar a la impresora que está seleccionada.
	nINIT	S	Activa en baja. Reinicializa la impresora. Reinicialización hardware.
Estado	nACK	E	Un pulso hacia abajo indica que el carácter fue recibido por la impresora.
	BUSY	E	En alta indica que la impresora está ocupada y no puede admitir más datos.
	PE	E	La impresora está sin papel (Paper Empty)
	SELECT	E	En alta indica que la impresora está lista (on line)
	nERROR	E	Indica que la impresora está en un estado de error.
Datos	DATA (1:8)	S	8 líneas de datos. En el modo SPP es únicamente de salida.

Tabla 3.7 Señales que intervienen en el Modo Centronics estándar.

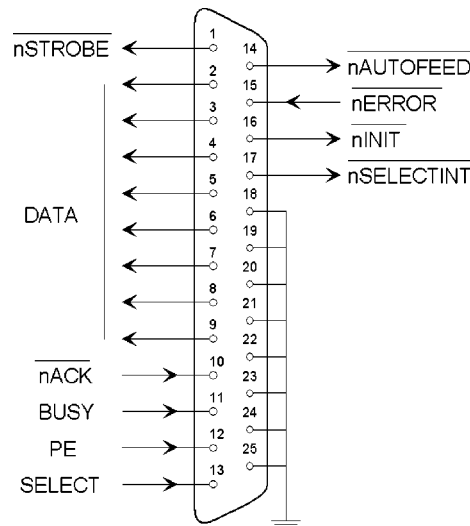


Fig. 3.21 Conector del interfaz Centronics convencional.
Las flechas indican el sentido de la información respecto del ordenador

3.10.2 Modo compatible (Centronics convencional)

El modo compatible, recibe también del nombre de puerto paralelo estándar (Standard Parallel Port ó SPP). En la Tabla (3.7) se muestran las distintas señales que intervienen, así como su significado y si son de entrada o salida al ordenador. En esta tabla se puede comprobar como la interfaz Centronics estaba muy orientada al manejo de impresoras, por los significados que tienen las señales de control y de estado. La descripción de estas señales y de las que aparecerán en las tablas correspondientes a los distintos modos se refieren a la fase de transmisión de datos, pero algunas de estas se utilizan para el cambio de modo o como señales de estado y control adicionales. En la figura (3.21) se muestra la distribución de estas señales sobre el conector.

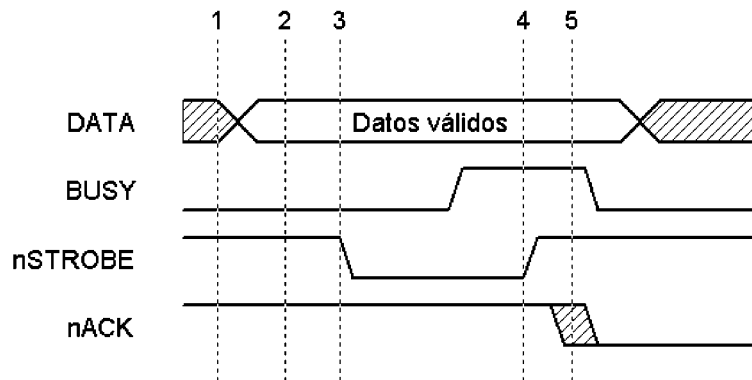


Fig. 3.22 Cronograma del ciclo de transferencia en el modo compatible

En la figura (3.22) se muestra el diagrama de tiempos de un ciclo de transferencia de datos en el modo compatible, sobre el que se pueden distinguir cinco fases:

- 1.- Escritura de los datos en el registro de datos de salida
- 2.- El programa lee el registro de estado, para ver si la impresora está ocupada (señal Busy).
- 3.- Si no está ocupada, escribe en el registro de control para dar la señal de disparo (STROBE)
- 4.- Nuevamente se escribe en el registro de control para cancelar la señal de disparo y se prepara para enviar un nuevo dato.
- 5.- La impresora reconoce la llegada del dato.

Como puede verse, el envío de un byte de datos requiere al menos cuatro instrucciones, esto hace que la velocidad de transferencia sea del orden de 150 Kbytes por segundo. Ha de tenerse en cuenta que al ser un protocolo con una gran componente software, será muy dependiente de la velocidad de la máquina. Esta velocidad es suficiente para comunicar con la mayoría de las impresoras de matriz de agujas y las de tecnología láser antiguas, pero no es suficiente para las impresoras láser de nueva generación y mucho menos para adaptadores de red local, discos extraíbles, etc. Aunque los ordenadores modernos pueden alcanzar velocidades muy superiores no deja de ser una fuerte restricción la dependencia de la velocidad de transmisión del tipo de procesador, además de sobrecargar la CPU con una tarea que debería realizarse de forma autónoma.

Algunos fabricantes han implementado un modo que usa una FIFO para transferir los datos en este modo a más alta velocidad. Este modo se conoce como Centronics rápido (Fast Centronics) o puerto paralelo con modo FIFO (Parallel Port FIFO mode). Cuando se emplea este modo, los datos se escriben en la FIFO y el hardware del controlador es el que se encarga de gestionar el protocolo aumentando la velocidad hasta unos 500Kbytes por segundo. Sin embargo, este modo no está contemplado en la norma IEEE-1284.

Señal SPP	Nombre en el modo Nibble	E/S	Descripción
nSTROBE	nSTROBE	S	No usado para la transferencia inversa
nAUTOFEED	HostBusy	S	En baja indica que el ordenador está listo para recibir un nibble. En alta indica que el nibble ha sido recibido
nSELECTIN	1284Active	S	En alta cuando está en un modo 1284
nINIT	nINIT	S	No usado para la transferencia inversa
nACK	PtrClk	E	En baja indica que hay un nibble válido. Se pone en alta en respuesta al flanco de subida de la señal HostBusy
BUSY	PtrBusy	E	Usado para el bit 3, 7
PE	AckDataReq	E	Usado para el bit 2, 6
SELECT	Xflag	E	Usado para el bit 1, 5
nERROR	nDataAvail	E	Usado para el bit 0, 4
DATA (1:8)	No se usan		

Tabla 3.8 mostrando las señales empleadas en el modo Nibble.

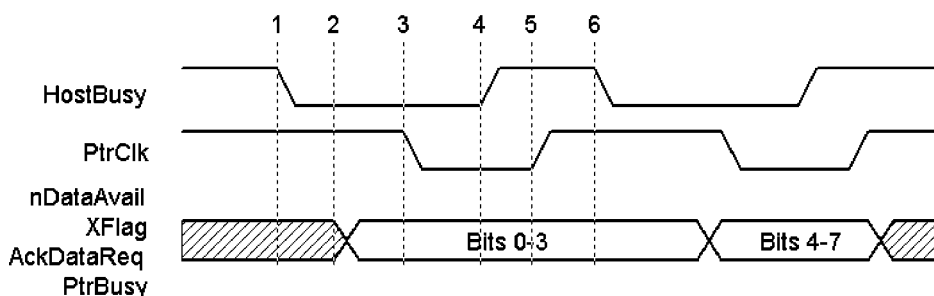


Fig. 3.23 Diagrama de tiempos durante una transferencia en el modo Nibble.

3.10.3 Modo nibble

Este modo es el más común para retornar datos desde la impresora al ordenador, e incluso para conectar dos ordenadores. Combinado con el modo compatible, permite una comunicación bidireccional sencilla. El estándar proporciona cinco líneas de información desde el periférico (normalmente una impresora) al ordenador que son usadas como señales de estado del dispositivo. Usando estas líneas, un periférico puede enviar un byte dividido en dos envíos de 4 bits (1 Nibble)

cada uno. El inconveniente de este modo, es que el ordenador debe formar nuevamente el byte leyendo dos veces en el registro de estado. La tabla (3.8) muestra las señales que intervienen en este modo y su correspondencia con las líneas del modo compatible (SPP).

En la figura (3.23) se muestran los puntos más relevantes durante la transferencia de un byte en el modo Nibble:

- 1.- El ordenador habilita la señal HostBusy para indicar que puede comenzar la transferencia.
- 2.- El periférico responde colocando el primer nibble en las líneas de estado.
- 3.- El periférico señala que los datos son válidos bajando la señal PtrClk
- 4.- El ordenador pone en alta la señal HostBusy para indicar que ha recibido el nibble e indica que todavía no está preparado para el siguiente.
- 5.- El dispositivo pone PtrClk en alta para indicar al host que hay un nuevo nibble preparado.
- 6.- Se repiten nuevamente las fases 1-5 para el segundo nibble.

El modo nibble, para la transmisión inversa (del periférico al ordenador) necesita muchas más instrucciones software que el modo compatible por lo que tiene una limitación de unos 50Kbytes por segundo de velocidad de transferencia. La principal ventaja de este modo es que no presupone la existencia de ninguna circuitería especial, y por lo tanto está disponible en todos los puertos paralelos Centronics. Este modo es útil en dispositivos que no requieren enviar muchos datos al ordenador como sucede con las impresoras, pero resulta inaceptable para adaptadores de red, CD-ROM, etc.

3.10.4 Modo byte

Al diseñar IBM su serie PS/2, modificó las etapas de excitación de las líneas de datos del puerto paralelo para permitir que pudiesen funcionar tanto para entrada como para salida de datos. Esto permite al dispositivo enviar un byte completo de datos empleando las mismas líneas por las que recibe los datos. Al poder enviar los bytes completos sin necesidad de partirlos en dos nibbles para enviarlos multiplexados en tiempo, la velocidad en modo inverso (del periférico al ordenador) se eleva considerablemente, igualándose a las velocidades de transferencia en sentido directo (del ordenador al periférico).

En la tabla (3.9) se muestran las señales que intervienen junto con una breve descripción de su significado, así como su correspondencia con las equivalentes al modo SPP y si son de entrada, salida o bidireccionales respecto al ordenador.

Señal SPP	Nombre en el modo Byte	E/S	Descripción
nSTROBE	HostClk	S	Es una señal de reconocimiento da un pulso bajo para indicar que el byte ha sido recibido
nAUTOFEED	HostBusy	S	En estado bajo indica que el ordenador está preparado para recibir un byte. Pasa a estado alto para indicar que el byte ha sido recibido
nSELECTIN	1284Active	S	En alto indica que está en un modo 1284
nINIT	nINIT	S	No usado. Debe estar en alta
nACK	PtrClk	E	En estado bajo indica que hay datos válidos en las líneas de datos. Pasa a alta en respuesta al flanco de subida de HostBusy
BUSY	PtrBusy	E	Línea de estado.
PE	AckDataReq	E	No usado
SELECT	Xflag	E	No usado en el modo Byte
nERROR	nDataAvail	E	Pasa a estado bajo para indicar que el byte está listo
DATA (1:8)	DATA	E/S	8 bits de datos del dispositivo al computador

Tabla 3.9 Señales que intervienen en el modo Byte.

La figura (3.24) muestra la transferencia en el modo byte, en la que se pueden distinguir los siguientes sucesos:

- 1.- El ordenador indica que puede recibir datos poniendo HostBusy en estado bajo
- 2.- El periférico responde colocando los datos en las líneas correspondientes
- 3.- El dispositivo indica que los datos son válidos mediante un pulso a través de PtrClk
- 4.- El ordenador pasa HostBusy a estado alto para indicar que ha recibido el dato y que todavía no está preparado para el siguiente envío.
- 5.- El periférico pasa CtrClk a estado alto como señal de reconocimiento al host y este responde bajando la señal HostClk

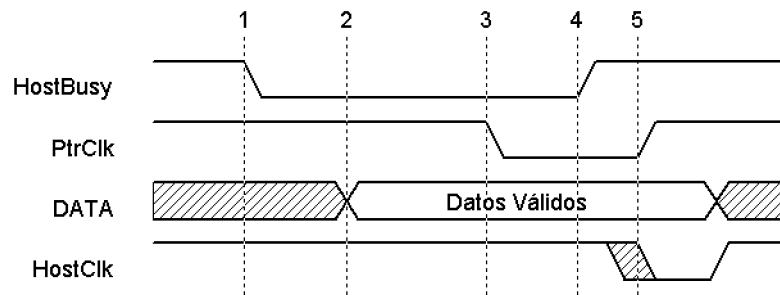


Fig. 3.24 Diagrama de tiempos durante una transferencia en el modo Byte

3.10.5 Modo EPP (Enhanced Parallel Port)

Este modo de funcionamiento fue desarrollado inicialmente por Intel, Xircom y Zenith Data Systems, para proporcionar un puerto paralelo de altas prestaciones y compatible con el Centronics convencional. Este protocolo fue implementado por Intel en el juego de circuitos integrados de soporte de la familia 386SL (integrado 82360 I/O).

El modo EPP ofrece grandes posibilidades sobre el interfaz estándar y fue adoptado rápidamente por numerosos fabricantes. Este protocolo, proporciona cuatro tipos de transferencia de datos:

- 1.- Ciclo de escritura de datos
- 2.- Ciclo de lectura de datos
- 3.- Ciclo de escritura de dirección
- 4.- Ciclo de lectura de dirección

Los ciclos de dirección están pensados para pasar direcciones, números de canal, comandos o información de control, y los de datos, para la transferencia de datos propiamente dicha. En la tabla se muestran las señales del modo EPP, y como en los modos anteriores, su correspondencia con el modo SPP.

Una de las características más sobresalientes de este modo es que la transferencia del byte está gestionada por el hardware del propio interfaz, con lo que el envío de un byte se reduce a una simple instrucción de salida. Esto permite elevar la velocidad de transferencia desde los 500Kbytes/s a los 2Mbytes por segundo. Los modos Nibble, Byte, EPP y ECP utilizan la técnica de protocolo con interbloqueo de forma que la transferencia se realiza siempre a la velocidad del elemento más lento, ya sea el periférico o el ordenador. Esto hace que tengamos una velocidad de transferencia adaptativa que resulta transparente tanto para el ordenador como para el dispositivo periférico.

Señal SPP	Nombre en el modo EPP	E/S	Descripción
nSTROBE	nWRITE	S	En baja indica una operación de escritura. En alta operación de lectura.
nAUTOFEED	nDATASTB	S	Activa en baja. Indica que una operación de lectura o escritura de datos está en proceso.
nSELECTIN	nADDRSTB	S	Activa en baja. Indica que una operación de lectura o escritura de dirección está en proceso.
nINIT	nRESET	S	Activa en baja. Reinicializa al dispositivo.
nACK	nINTR	E	Interrupción. Utilizada por el periférico para producir una interrupción en el ordenador y solicitar así su atención.
BUSY	nWAIT	E	En baja indica que esta listo para comenzar un ciclo. En alta indica que está listo para terminar.
PE	Definidas por el usuario	E	Estas tres señales pueden ser utilizadas de forma diferente por cada periférico.
SELECT		E	
nERROR		E	
DATA (1:8)	AD	E/S	Para enviar o recibir tanto los datos propiamente dichos como las direcciones.

Tabla 3.10 Señales del protocolo EPP

Esta característica de interbloqueo hace referencia a que cada una de las señales de control es reconocida por el lado opuesto del interfaz. Esto se ve más claro si nos fijamos en la transferencia de un dato en el modo EPP, como se muestra en la figura (3.25).

- 1.- El programa ejecuta una instrucción de salida, representada por una señal *IOW* en el bus del sistema.
- 2.- La línea *nWrite* pasa a estado bajo para indicar que se trata de una escritura, y los datos pasan a la salida.
- 3.- Se da la señal de disparo (bajando *nDataStrobe*) para indicar que los datos han sido colocados, y permanecerá así hasta que *nWAIT* pase a estado bajo.
- 4.- El puerto del ordenador (no un programa) espera que el dispositivo cambie el estado de *nWAIT*.
- 5.- La señal de disparo es retirada (pasa nuevamente a alta) y el ciclo termina.
- 6.- Finaliza el ciclo de ejecución de la instrucción de salida.
- 7.- *nWAIT* vuelve a estado bajo para indicar que el siguiente ciclo puede comenzar.

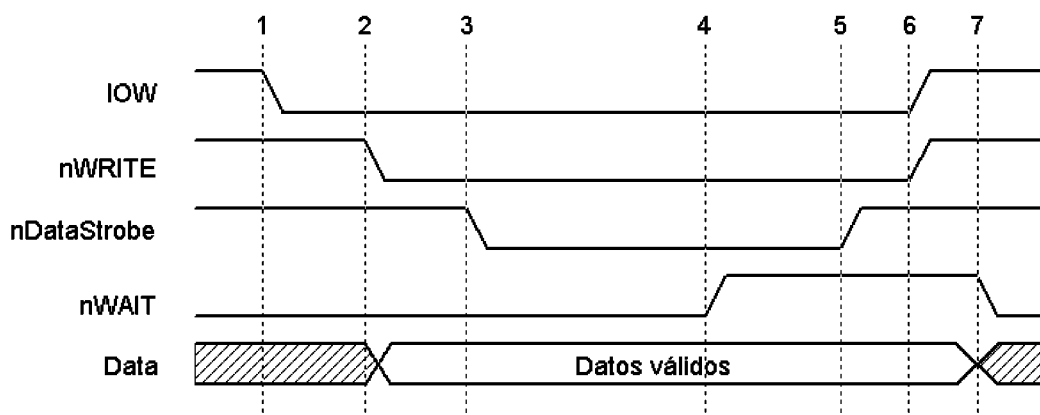


Fig. 3.25 Diagrama de tiempos durante una operación de escritura de datos en modo EPP

De esta manera el periférico puede controlar el tiempo de set-up (tiempo entre 3 y 4) que necesita para su correcto funcionamiento. Esta característica de interbloqueo, también hace que la transferencia sea independiente de la longitud del cable. Sin embargo, la longitud del cable puede hacer que la comunicación se haga lenta debido a que los flancos de subida o bajada queden muy suaves y también que se produzcan errores debido a que la contaminación con ruido será mayor cuanto mayor sea la longitud del cable. La tabla (3.10) muestra las señales en este modo.

El diagrama de tiempos para un ciclo de lectura, sería idéntico pero con la señal nWRITE en estado alto en lugar de bajo. Para los modos de escritura y lectura de direcciones, tendríamos unos diagramas de tiempo idénticos con la única salvedad de que ahora la señal de disparo la dará la señal nADDRSTB en lugar de nDATASTB. Con el modo EPP se proporcionan nuevos registros, que permiten incluso transferencias de 16 o 32 bits. En la tabla (3.11) se muestran estos registros.

Nombre del puerto	Offset	Modo	R/W	Descripción
SPP Data Port	0	SPP/EPP	W	Puerto de salida de datos estándar. Sin auto-disparo -> necesita dar el disparo escribiendo en el registro de control.
SPP Status Port	1	SPP/EPP	R	Registro de estado
SPP Control Port	2	SPP/EPP	W	Registro de control del protocolo
EPP Address Port	3	EPP	R/W	Realiza una lectura o escritura de dirección con interbloqueo (auto-disparo)
EPP Data Port	4	EPP	R/W	Realiza una lectura o escritura de dato con interbloqueo (auto-disparo)
No definido	5-7	EPP	R/W	Los emplean algunas implementaciones para permitir transferencias de 16 o 32 bits

Tabla 3.11 Registros que utiliza el protocolo EPP.

Se muestran también cuales son compartidos con el modo Centronics convencional o SPP.

3.10.6 Modo ECP (Extended Capability Port)

Este protocolo fue propuesto por Hewlett Packard y Microsoft como un modo avanzado para la comunicación con impresoras gráficas y escáner fundamentalmente. Una característica de estos dos tipos de dispositivos, es que normalmente precisan la transferencia de grandes cantidades de datos entre los que hay una gran redundancia. Este hecho sugiere el empleo de alguna técnica de compresión. El método escogido es el RLE (Run Length Encoding). Otras características que incorpora este protocolo es una FIFO en cada extremo del interfaz y acceso DMA.

Este modo también incorpora direccionamiento, aunque es conceptualmente diferente al del modo ECP. Este modo está orientado a manejar varias unidades lógicas dentro de un mismo dispositivo. Esto permite que se tengan un fax, una impresora y un módem conectados como un único dispositivo y se direccionen lógicamente. De esta forma, se pueden recibir datos del módem mientras la impresora está procesando datos. En el modo compatible, si la impresora está ocupada, activa la señal de BUSY y el interfaz queda paralizado, sin embargo con el direccionamiento, basta con direccionar otro dispositivo para realizar alguna transferencia con él y esperar a más adelante para continuar con el trabajo de impresión. Hay que tener en cuenta que estos distintos periféricos constituyen un solo equipo que se conecta mediante un único cable a la unidad central.

Como en el resto de modos IEEE-1284, el protocolo ECP redefine las señales SPP dándoles un nuevo significado. Al contrario que con los otros modos, cuando se propuso el estándar para este modo no sólo se proponía el significado e interpretación de las señales, sino también varios registros para permitir la comunicación. En la tabla (3.12) se muestran las señales en el modo ECP y en la tabla (3.13) los registros que emplea.

Señal SPP	Nombre en el modo ECP	E/S	Descripción
nSTROBE	HostClk	S	Utilizado con PeriphAck para transferir datos en sentido directo.
nAUTOFEED	HostAck	S	Proporciona el estado comando/dato en sentido directo.
nSELECTIN	1284Active	S	En alta indica que se trabaja en algún modo IEEE-1284
nINIT	nReverseReq	S	Pasa a estado bajo para indicar el sentido inverso.
nACK	PeriphClk	E	Utilizado con HostAck para transferir datos en sentido inverso.
BUSY	PeriphAck	E	Utilizado con HostClk para transferir datos o direcciones en sentido directo. Proporciona el estado comando/dato en sentido inverso.
PE	nAckReverse	E	Pasa a baja como reconocimiento a nReverseReq
SELECT	Xflag	E	Bandera de extensión.
nERROR	nPeriphReq	E	En estado bajo indica que un dato está listo para ser enviado del periférico al ordenador.
DATA (1:8)	Data	E-S	Datos bidireccionales.

Tabla 3.12 Señales en el modo ECP.

El protocolo ECP proporciona dos tipos de transferencia de información, tanto en sentido directo, como inverso, que son el ciclo de datos y el de comandos. En la figura (3.26) se muestra el diagrama de tiempos correspondiente a dos ciclos de transferencia en sentido directo (uno de datos y otro de comando) con los siguientes eventos:

- 1.- El ordenador coloca un byte sobre las líneas de datos, y señala que se trata de un dato (en lugar de un comando) poniendo en alta la señal HostAck.
- 2.- El ordenador pasa a baja la señal HostClk para indicar que el dato es válido.
- 3.- El periférico envía el reconocimiento al ordenador poniendo PeriphAck en alta.
- 4.- El ordenador pone en alta HostClk. Este flanco debería ser usado por el periférico para tomar el dato.
- 5.- El periférico pone PeriphAck en baja para indicar que está preparado para el siguiente byte.
- 6.- El ciclo se repite, pero ahora el byte enviado corresponde a un comando, ya que HostAck está en baja.

Nombre	R/W	Modo ECP	Función
Data	R/W	000-001	Registro de datos
ecpAfifo	R/W	011	Dirección de la FIFO ECP
dst	R/W	todos	Registro de estado
dcr	R/W	todos	Registro de control
cfifo	R/W	010	FIFO del puerto paralelo
ecpDfifo	R/W	011	FIFO de datos ECP
tfifo	R/W	110	Test FIFO
cnfgA	R	111	Registro de configuración A
cnfgB	R/W	111	Registro de configuración B
ecr	R/W	todos	Registro de control ampliado

Tabla 3.13 Registros empleados en el modo ECP

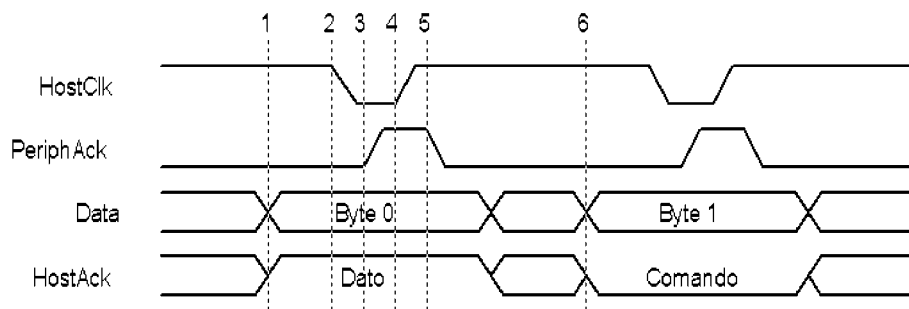


Fig. 3.26 Envío de un dato y luego un comando en sentido directo.

En el protocolo EPP el software de control puede enviar o recibir datos desde el periférico, sin mayor problema, sin embargo, en el modo ECP el cambio de dirección debe ser negociado. Esto se muestra en la figura (3.27) donde puede verse un diagrama de tiempos similar, pero ahora corresponde a un envío en sentido inverso. Los puntos relevantes de esta transferencia son:

- 1.- El ordenador solicita invertir el canal de comunicaciones poniendo nReverseReq en estado bajo.
- 2.- El periférico contesta que está listo para la comunicación inversa bajando nAckReverse.
- 3.- El periférico coloca el byte sobre las líneas de datos e indica que se trata de un dato (no un comando) poniendo PeriphAck en estado alto.
- 4.- El periférico pone PeriphClk en baja para indicar un dato válido.
- 5.- El ordenador envía la contestación de reconocimiento poniendo HostAck en alta.
- 6.- El periférico pone PeriphClk en alta. Este flanco debe ser usado por el ordenador para tomar el dato.
- 7.- El ordenador pone HostAck en baja para indicar que está preparado para el siguiente byte.
- 8.- El ciclo se repite, pero en este caso se trata de enviar un comando porque PeriphAck esta en estado bajo.

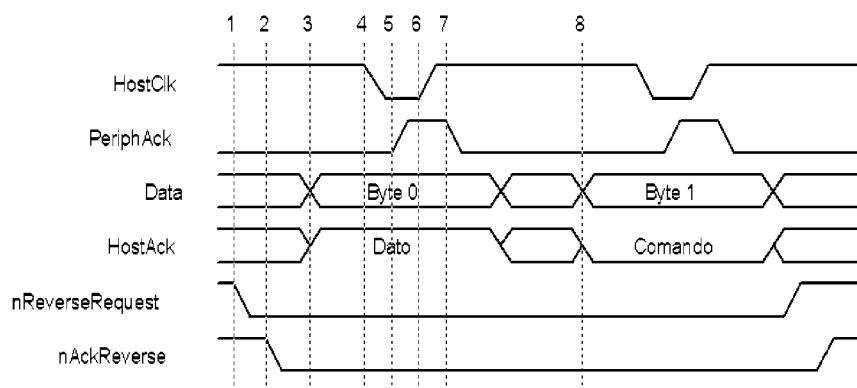


Fig. 3.27 Envío de un dato y luego un comando en sentido inverso.

3.10.7 Negociación de modo

Hasta ahora hemos descrito los distintos modos del interfaz IEEE-1284. Los periféricos no tienen porque tener implementados todos los modos anteriores, con lo que se hace necesario un método para determinar cuales son las posibilidades del dispositivo conectado al puerto y que permita al ordenador establecer el modo apropiado de funcionamiento. Para solventar el problema, se introdujo el concepto de negociación de modo. Mediante el proceso de negociación, un ordenador establece comunicación con el periférico para conocer los modos que éste implementa y elegir uno de ellos.

La negociación es una secuencia de eventos que debe realizarse a través del interfaz, entre el ordenador y el periférico, pero que no debe tener efecto sobre un periférico antiguo, ajeno a los nuevos modos del estándar. Es decir, un dispositivo más viejo que solo soporta el modo compatible, correspondiente al modo Centronics convencional, no responderá al proceso de negociación.

El byte de extensión se utiliza durante la negociación para que el periférico entre en un modo determinado. La tabla (3.14) muestra los valores permitidos para este byte. La señal Xflag es utilizada por el periférico para indicar que el modo solicitado está disponible. Esta señal estará el estado alto para todos los modos, salvo para el modo Nibble que está presente, como ya se señaló, en todos los dispositivos, incluidos los más antiguos. El bit de enlace (Request Extensibility Link) se utiliza como una forma de contemplar posibles ampliaciones futuras y no se usa.

Bit	Descripción	Valores válidos (7654 3210)
7	Request Extensibility Link	1000 0000
6	Request EPP Mode	0100 0000
5	Request ECP Mode with RLE	0011 0000
4	Request ECP Mode without RLE	0001 0000
3	Reservado	0000 1000
2	Request Device ID	Modo de retorno de datos: Nibble Mode 0000 0100 Byte Mode 0000 0101 ECP Mode without RLE 0001 0100 ECP Mode with RLE 0011 0100
1	Reservado	0000 0010
0	Byte Mode	0000 0001
ninguno	Nibble Mode	0000 0000

Tabla 3.14. Valores del byte de extensión.

- 1.- El ordenador coloca en las líneas de datos el byte de extensión para solicitar un determinado modo.
- 2.- Una vez hecho esto, pone nSelectIn en alta y nAutoFeed en baja para indicar que comienza una secuencia de negociación. Recuérdese que en el puerto Centronics convencional nSelectIn en alta significa que la impresora no está seleccionada, con lo que un dispositivo antiguo no se dará por aludido y no responderá a los siguientes eventos.
- 3.- Un periférico IEEE-1284 responderá poniendo nAck en estado bajo, y nError, PE y Select en alto. Un dispositivo que no sea IEEE-1284 no responderá.
- 4.- El ordenador establece nStrobe en estado bajo como señal de disparo indicando al dispositivo que el byte de extensión está disponible sobre las líneas de datos.
- 5.- El ordenador ahora, sube a estado alto tanto nStrobe como nAutoFeed para indicar al periférico que lo ha reconocido como dispositivo IEEE-1284.
- 6.- El periférico responde bajando PE; Y pone nError en estado bajo si el periférico dispone de canal inverso, e indica que el modo no está disponible poniendo Select en estado bajo.
- 7.- Por último el periférico pone nAck en estado alto para indicar que la secuencia de negociación ha terminado y que todas las señales están en el estado solicitado, si éste es soportado.

En la figura (3.28) se muestra el diagrama de tiempos de la fase de negociación:

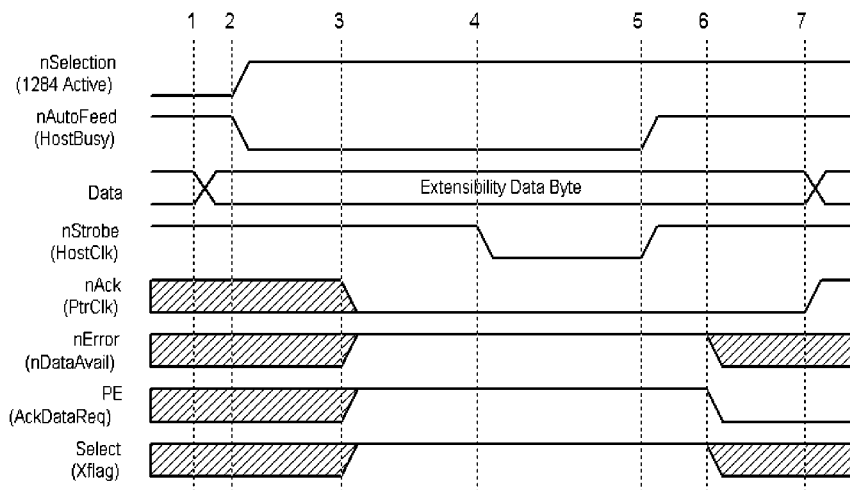


Fig. 3.28. Diagrama de tiempos correspondiente a una negociación de modo IEEE-1284.

3.11 BUS IEEE-488

La interfaz IEEE-488 es el resultado de la normalización de un bus propietario de la compañía Hewlett-Packard. Esta empresa comenzó el desarrollo del bus en 1965, para la interconexión de los instrumentos de laboratorio de la misma. El objeto era un bus de propósito general, destinado a simplificar el diseño y la integración de equipos de medida entre sí y especialmente de estos con el ordenador. Dicha simplificación se consigue al reducir al mínimo los problemas tanto eléctricos como mecánicos y de compatibilidad funcional entre equipos, poseyendo la suficiente flexibilidad para acomodar un amplio y creciente número de productos. El nombre inicial fué el de HPIB (Hewlett Packard Interface Bus). Rápidamente, numerosas empresas empezaron a comercializar equipos que incorporaban este tipo de interfaz con el nombre más habitual de GPIB (General Purpose Interface Bus). Estos trabajos iniciales fueron del interés de la comisión Electrotécnica Internacional (IEC) y del Instituto de Ingenieros Eléctricos y Electrónicos (IEEE) y respaldaron el borrador propuesto por Hewlett-Packard. El IEC-625 y el IEEE-488/1978 son los nombres oficiales de las normas publicadas por los dos organismos anteriormente mencionados.

La interfaz IEEE-488 se aplica a sistemas de interconexión de instrumentos en los cuales:

- El intercambio de información entre los equipos interconectados, sea de naturaleza digital. Este aspecto parece obvio, puesto que todos los periféricos que hemos tratado hasta ahora han sido de naturaleza digital. Sin embargo, este bus se concibió para interconectar instrumentos de medida, como generadores de señal, osciloscopios, analizadores de espectro, fuentes de alimentación, multímetros, etc. que en principio tienen una naturaleza analógica.
- El número de equipos a interconectar no exceda de 15. La limitación a 15 aparatos se debe a que por ser un procedimiento asíncrono la sobrecarga del bus hace que el funcionamiento no sea fiable para una carga mayor.
- Las longitudes totales de transmisión sobre los cables de interconexión no exceda de 20 metros o de dos metros por equipo, cuando no se utilicen técnicas especiales de ampliación del bus.
- La velocidad de los datos en la interconexión y en cualquier línea de la misma no supere la cantidad de 1 MByte/segundo. Consideraciones prácticas hacen que el límite de velocidad de transmisión de datos sea del orden de 250 KBytes/s. No obstante, la revisión IEEE-488.2

garantiza velocidades de 1MB/segundo y casi cualquier interfaz comercial supera con creces esta velocidad.

Las normas IEEE-488 e IEC-625 son totalmente compatibles a nivel funcional y eléctrico, pero no a nivel mecánico. La propuesta mecánica de IEC contempla la utilización de un conector de 25 contactos, exactamente igual al utilizado por los interfaces RS-232 (CCITT V.24) mientras que la norma de IEEE propone un conector de 24 contactos tipo Ribbon. Sin embargo, esta diferencia mecánica es fácil de subsanar utilizando los adaptadores adecuados. La principal ventaja de los conectores Ribbon es que incorporan un macho y una hembra de tal forma que por el un lado se conecta a un determinado instrumento u ordenador y por el otro lado se puede conectar un nuevo cable que enlace al siguiente equipo.

3.11.1 Estructura del bus

El sistema de interface IEEE-488 utiliza una estructura de bus de línea compartida, es decir los equipos comparten las líneas de señal. La estructura del bus consiste en 16 líneas de señal (ocho de datos y ocho de control), y ocho líneas de masa en una configuración paralela y manteniendo un flujo ordenado de información entre equipos e interconexión. Cualquier equipo conectado al GPIB puede ejecutar una o más de las siguientes funciones:

- A) 'Talker' = Locutor. Equipo capaz de transmitir datos si es direccionado. Aunque más de un equipo pueda tener esta funcionalidad, en cada momento sólo puede haber un locutor activo conectado al bus.
- B) 'Listener' = Oyente. Instrumento direccionado para recepción de datos. Varios escuchas pueden estar activos sobre un mismo interfaz simultáneamente. Esto permite enviar datos a varios dispositivos simultáneamente y en este caso la transferencia será a la velocidad del equipo más lento.
- C) 'Controller' = Controlador. Unidad destinada a direccionar los instrumentos conectados al interfaz, En la mayoría de los casos será un ordenador. Asimismo puede definir una unidad como hablador o como escucha si aquella es funcionalmente ambivalente. Lógicamente el controlador puede asumir también funciones de habla y escucha. La única restricción es que sólo puede existir un controlador activo, de forma simultánea en el bus, de forma que en sistemas multimaestros, uno de los controladores toma el bus, mientras los demás permanecen pasivos o adoptan la apariencia funcional de escuchas o habladores.

La figura (3.29) muestra la estructura de conexión de diferentes equipos entre sí por medio de este bus.

En general la denominación de controlador es atribuible a uno o varios instrumentos con capacidades de control sobre todo el sistema de medida, es decir las funciones propias de cada equipo locutor u oyente y las funciones propias del interfaz. Generalmente, el controlador toma la forma de un ordenador, aunque esto no es imprescindible.

Desde el punto de vista del bus, el controlador asume el pilotaje de todas las funciones propias del mismo, cuidando que las transferencias de datos a su través se efectúen correctamente. Otra función básica del controlador consiste en determinar qué instrumentos actúan como locutores, cuales como oyentes y en que momento lo hace cada uno.

En un sistema pueden coexistir diversos controladores, pero solamente uno de ellos tendrá oficio de controlador general del sistema, ya que la especificación del interfaz no permite la existencia de más de un controlador maestro. Únicamente el controlador central del sistema puede

activar los circuitos de validación con la señal REN (Remote ENable) y de invalidación IFC (Interface Clear) del interfaz.

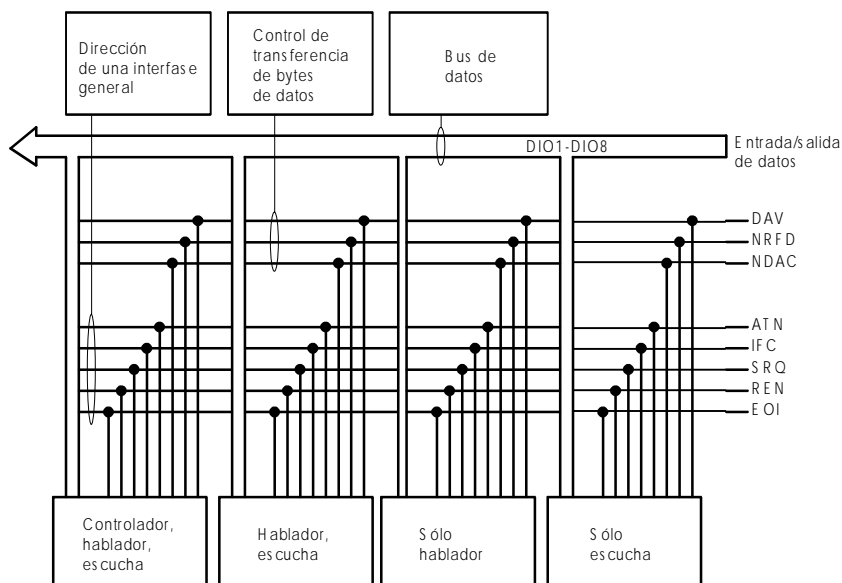


Fig. 3.29 Estructura de una conexión por GPIB

Para sistemas pequeños, en los que no se requieran posibilidades de reconfiguración dinámica de equipos o en aquellos otros donde no existe procesamiento de las señales medidas, es posible prescindir del controlador, siempre que se configuren las funciones de los instrumentos de forma manual. Una solución de este tipo es evidentemente rígida y solamente será válida en sistemas extremadamente simples. Un ejemplo de esta situación se presenta, cuando se tiene por ejemplo un osciloscopio y queremos imprimir la pantalla en una impresora o plotter. A la impresora, normalmente no hay que hacerle nada, puesto que siempre está configurada como "listen only" (Sólo escucha); sin embargo un osciloscopio moderno, puede realizar cualquier función. Por eso, cuando vayamos a imprimir deberemos configurar el osciloscopio como "talk only" (Sólo habla). Esto se puede hacer normalmente desde el panel de control del propio instrumento, o a través de los menús de pantalla si el equipo incorpora pantalla táctil.

3.11.2 Examen funcional del bus

Todas las líneas de señal del bus de la interfaz funcionan en lógica negativa y niveles TTL y están cargadas por circuitos en colector abierto (voltaje menor de 0.8 V es un '1' y superior a 2.5 V es un '0'). Estas líneas se pueden clasificar en tres grupos bien diferenciados (ver figura 3.30)

FUNCIONES DEL EQUIPO DE MEDIDA	FUNCIONES INTERFAZ	DATOS ENTRADA/SALIDA					
		DAV	NRFD	NDAC	REN		
		CONTROL DE TRANSFERENCIAS					
		COMANDOS GENERALES					
		BUS DE DATOS					

Fig. 3.30 Líneas del bus

a) Líneas de datos

Existen ocho líneas bidireccionales (DIO1 - DIO8) que se utilizan para la transferencia de datos entre un equipo que los envía ('talker') y tantos otros como estén en ese momento recibiendo ('listeners'). Normalmente se utiliza un código ASCII normalizado de 7 bits, con el octavo disponible para el control de paridad. La información transferida incluye los comandos de control de la interface, direcciones y datos dependientes de los equipos. Los equipos deben poseer registros de almacenamiento de lectura a fin de asegurar su recepción.

Las ocho líneas bidireccionales son utilizadas para:

- las medidas
- las instrucciones de programación de equipos
- las direcciones
- las palabras de estado
- comandos universales multilínea

b) Líneas de control

Existen ocho líneas de control, de las que tres (DAV, NRFD y NDAC) son líneas de protocolo y se usan para coordinar el intercambio de información entre los equipos e instrumentos conectados al interfaz. El objetivo de las líneas de protocolo es facilitar el manejo del bus por el controlador, así como permitir una gran flexibilidad en la conexión de equipos e instrumentos de diversa índole. Gracias a las líneas de protocolo la transferencia de datos puede ser asíncrona y la velocidad de transferencia puede ajustarse automáticamente a la velocidad del equipo activo más lento. Cabe asimismo la posibilidad de que se produzca aceptación de datos por más de un equipo de forma simultánea. Por otra parte, cada una de las cinco líneas restantes presenta una función específica entre el controlador y el resto de equipos conectados al sistema.

- | | |
|------|--|
| DAV | DATA VALID (Dato válido). Es una de las tres líneas de control de la transmisión de datos. Un '1' (< 0.8V) indica que el dato está disponible en las líneas de datos. Se controla por el locutor activo o por el controlador |
| NRFD | NOT READY FOR DATA (No preparado para recibir datos). Es otra de las tres líneas de control mencionadas anteriormente. Un '0' indica que los equipos que reciben 'listeners' están dispuestos a admitir los datos. Es controlada por todos los escuchas activos o por aquel equipo que esté recibiendo los comandos del bus. |
| NDAC | NOT DATA ACCEPTED (Dato no aceptado). Es la tercera de las líneas mencionadas anteriormente, indica al equipo emisor que los datos han sido leídos. Un '0' indica que todos los receptores han leído los datos. Esta línea viene controlada por los escuchas activos o por todos los dispositivos que estén recibiendo comandos del interfaz. |
| ATN | ATTENTION (Atención). Esta señal es utilizada por el controlador del bus para indicar que se va a mandar una orden. Un '1' en ATN indica que los datos enviados son órdenes. Todos los equipos deben monitorizar esta señal de forma continua y responder a ella antes de 200 ns. Cuando esta línea es activa, la señal ATN coloca al bus en modo comando, de modo que todos los equipos acepten datos y los interpreten como comandos. Todos los equipos son receptores de comandos. Cuando la línea está desactivada, la señal ATN coloca al interfaz en modo dato. En este modo un hablador activo proporciona datos solo a los escuchas activos, el resto de escuchas ignoran los datos. |

IFC	INTERFACE CLEAR (Liberar interfaz). Esta señal se usa para inicializar todos los aparatos del bus, es decir situarlos en un estado no direccionado y no activo, a fin de conseguir una situación transparente de los mismos antes de iniciar una nueva secuencia de operaciones en el bus. Todos los equipos deben monitorizar constantemente esta señal y responder a ella antes de 100 μ s. Un '1' causa la vuelta a las condiciones iniciales de todos los aparatos conectados al bus.
REN	REMOTE ENABLE (Permiso remoto). Esta línea es usada por el controlador del sistema para disponer los equipos conectados al bus en el modo de programación remota. Cuando la línea es activa, todos los escuchas se colocan en operación remota cuando se les direcciona como tales. Cuando la línea está desactivada, todos los equipos vuelven al modo local. Cualquier equipo capaz de operar de forma remota y local debe necesariamente monitorizar en todo momento la línea REN y debe ser capaz de responder a un cambio de nivel en la misma antes de 100 μ s. Un '0' permite a todos los aparatos del bus ser controlados por el bus GPIB.
SRQ	SERVICE REQUEST (Solicitud o Demanda de servicio). Esta señal se usa para que los equipos que requieren servicio se lo indiquen al bus (p. ej. han completado una tarea, se ha producido un error, etc.). Cuando el controlador detecta un '0' en SRQ muestrea los aparatos en búsqueda del que está requiriendo sus servicios para atenderlo a continuación.
EOI	END OF IDENTIFY (Final de identificación) Esta señal tiene dos funciones. Cualquier equipo puede poner un '0' en EOI indicando que ha terminado la transmisión. El controlador puede usar EOI para iniciar un muestreo en paralelo. (Cuando se envían juntos un ATN y un EOI, los aparatos conectados al bus presentan sus bits de estado en las líneas de datos).

Aclaremos ahora algunos conceptos introducidos en la descripción de las señales precedentes. En primer lugar, hay que tener en cuenta que las salidas de los equipos son en colector abierto, lo que permite realizar una OR cableada. Esto resulta especialmente interesante cuando un equipo emite y varios reciben. La línea que indica que los datos han sido recibidos sólo pasará a estado alto cuando todos los dispositivos pongan su salida en alto, con lo que cuando el emisor detecta un estado alto en esta línea de reconocimiento tiene la seguridad de que todos los dispositivos oyentes han aceptado el dato. Ver figura (3.31).

El segundo aspecto a tener en cuenta es el de modo local o remoto. La mayoría de equipos diseñados para ser conectados a través del bus GPIB son relativamente complejos y pueden funcionar de forma autónoma y ser controlados manualmente a través de su panel de control. Pensemos por ejemplo en un osciloscopio, un generador de señal o una fuente de alimentación digital. Estos equipos pueden desarrollar toda su funcionalidad sin necesidad de ser conectados a otros equipos, al contrario de lo que sucede con un disco duro, o una impresora, que no tienen ninguna utilidad si no los conectamos a un ordenador. Sin embargo, para aumentar sus posibilidades, puede que deseemos conectar esos equipos entre sí o a un ordenador central. Cuando varios de estos equipos son conectados al bus, podemos controlarlos a través de su panel frontal (modo local) o a través del bus (modo remoto). Habrá situaciones en las que queramos que el control remoto (realizado desde un ordenador a través del bus) no se vea interferido por manipulaciones del panel de control del equipo, para lo que resultará conveniente bloquear este modo de funcionamiento. Esto se consigue, como se comentará más adelante, con el comando: 'Local Lockout' (Bloqueo del Modo Local).

3.11.3 Protocolo de operación

Los datos se transmiten en el GPIB byte a byte. Las líneas de control DAV, NRFD y NDAC manejan la transferencia de bytes realizando un intercambio entre los equipos transmisor y receptor(es). El procedimiento de intercambio asegura que un byte no es enviado hasta que todos los receptores están preparados, que cada receptor sólo lee el bus cuando el byte válido está allí, y que el emisor mantiene su dato en el bus hasta que ha sido leído por todos los receptores. El diagrama de tiempos de la figura (3.31) ilustra como se usan las señales DAV, NRFD y NDAC para realizar esto.

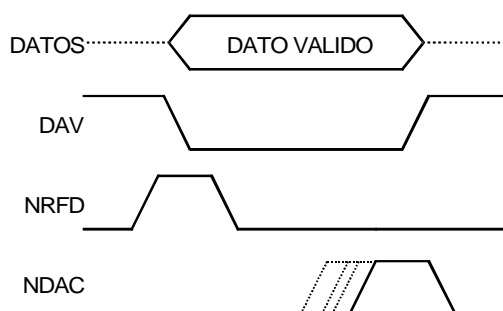


Fig. 3.31 Diagrama de tiempos de las líneas de control del GPIB. En la última de las curvas correspondiente a la señal NDAC se muestra el efecto de la OR cableada sobre el bus. Las líneas punteadas muestran cuando los distintos equipos reconocen la aceptación del dato, pero la línea no pasa a estado alto hasta que todos ellos lo han hecho.

Antes de poner un dato en el bus el equipo transmisor debe esperar a que la señal NRFD se ponga a '0' (> 2.5 V). Como está conectada en colector abierto, esto quiere decir que todos los aparatos del bus están dispuestos para recibir datos, es decir, que ninguno va a utilizar el bus para transmisión. Una vez la línea NRFD está en alto, el emisor puede poner sus datos en el bus. Además pone la señal DAV a bajo nivel (dato válido en el bus). Cuando el receptor detecta un nivel bajo en DAV, lee el dato del bus. Cuando cada receptor termina de leer el dato libera NDAC. Cuando el último libera NDAC, ésta va a '0' (nivel alto), señalando al emisor que el dato ha sido aceptado por todos los receptores. Una vez que ha sido aceptado por todos los aparatos del bus, el emisor puede quitar el dato del bus y liberar DAV.

Los equipos conectados al bus pueden enviar datos, recibirlos o controlar el bus o cualquier combinación de estas funciones. El controlador debe fijar qué equipo es el emisor y cuales los receptores, además el controlador ejecuta otras funciones de control del bus. Esto lo realiza por medio de las órdenes del bus, que se envían de forma análoga a los datos pero con la señal ATN puesta a '1'. Las órdenes son leídas por todos los equipos.

El controlador puede enviar cuatro tipos de órdenes o comandos: direcciones ('adress'), escuche ('listen'), hable ('talk') y universal ('universal'). Para dar estas órdenes sólo se usan los siete primeros bits del bus de datos. El tipo de comando se da en los bits 5, 6 y 7 (tabla 3.15). Existen también comandos secundarios que se usan para configurar un equipo para un muestreo paralelo o enviar direcciones secundarias.

b7	b6	b5	Tipo de comando
0	0	0	De dirección
0	0	1	Universal
0	1	x	De habla
1	0	x	De escucha
1	1	x	Secundario

Tabla 3.15 Código del tipo comando del GPIB

Comandos hable y escuche

Estos comandos sirven para indicar el paso al estado de emisor o receptor de un determinado equipo. Los últimos cinco bits permiten direccionar el equipo a que nos referimos, éstos pueden tener bien por hardware o software asignadas direcciones en el bus de 0 a 30 (la dirección 31 se usa en los comandos UTN ('untalk') y UNL ('unlisten') que mandan a todos los equipos a estado de reposo.

Comandos universales

Son los que afectan a todos los equipos del bus. Existen cinco tipos de comandos universales:

- LLO 'Local Lockout' (11H) (bloqueo del modo local). Esta orden inutiliza los controles manuales de los equipos a fin de que no exista conflicto entre las instrucciones enviadas por el bus y los mandos locales de los equipos.
- DCL 'Device Clear' (14H) (inicialización de todos los equipos). Esta orden reinicializa todos los equipos en el bus.
- PPU 'Parallel Poll Unconfigure' (15H). Esta orden resetea las respuestas a un muestreo o encuesta paralelo, permitiendo realizar un nuevo muestreo.
- SPE 'Serial Poll Enable' (18H) y 'Serial Poll Disable' (19H). Estas órdenes se usan para ejecutar un muestreo serie de los equipos del bus. Cuando el comando detecta una demanda de servicio, debe determinar qué equipo es el que lo demanda antes de actuar. El controlador envía un SPE y a continuación manda un comando 'hable' a cada equipo (uno detrás de otro) y éstos contestan con un byte de estado. Cuando encuentra el equipo que quiere emitir, el controlador envía un SPD que causa la vuelta al estado normal de todos los equipos y así el controlador puede pasar a atender el servicio requerido.

Comandos de direcciones

Estas órdenes sólo afectan a equipos a los que previamente se les ha enviado una orden de 'escuche' y por ello sólo afectan a ciertos equipos. Existen cinco comandos de direcciones:

- GTL 'Go To Local' (001). Esta orden cancela el comando universal LLO volviendo el control de los equipos que están recibiendo a modo local.
- SDC 'Selected Device Clear' (004). Esta orden libera todos los equipos que han recibido la orden 'escuche'.
- GET 'Group Trigger' (008). Esta orden sincroniza la operación de un cierto número de equipos. Estos han de ser previamente programados en su actuación, y cuando reciben un GET comienzan simultáneamente sus tareas.
- TCT 'Take Control' (009). Esta orden transfiere el control del bus de un controlador a otro, que ha sido previamente puesto en condición de recibir. Tras recibir esta orden toma el control del bus el nuevo controlador y comienza a enviar órdenes.

PPC 'Parallel Poll Configure' (005). Esta orden prepara a cualquier equipo para participar en un muestreo paralelo. Un muestreo paralelo se efectúa para conocer qué equipo está requiriendo servicio cuando SQR se pone a '0'.

El muestreo paralelo se inicia cuando el controlador envía un ATN y un EOI simultáneamente; esto causa que los equipos que han sido configurados para un muestreo paralelo pongan un bit de estado en una de las líneas del bus. El controlador a continuación examinará el bus para determinar qué equipo requiere servicio. El comando PPC se usa para especificar qué bit usará el equipo para especificar su estado. Después del PPC, el controlador envía al equipo un comando secundario cuyos tres bits menos significativos indican el bit que será usado para especificar su estado (p. ej. 010 indica que será el bit tercero), el cuarto bit señalará cómo se indicará el estado del equipo (si es '0' el bit de estado se pondrá a '0' para indicar que se desea servicio, y si es '1' se pondrá a '1').

Durante un muestreo paralelo más de un equipo puede usar el mismo bit para indicar su estado. Si este bit es '0' en los dos, la línea en el bus será '0' si uno de ellos requiere servicio. Si fuera '1' la línea del bus será '1' si los dos equipos requieren servicio.

Implementación del interfaz GPIB

Existen numerosos fabricantes que suministran tarjetas de interfaz GPIB para casi cualquier tipo de sistema. Suelen estar basadas en circuitos integrados VLSI GPIB, por ejemplo los Intel 8291 talker/listener, 8296 GPIB controller, y un par de 8293 GPIB transceivers. También Texas Instruments (TMS9914), Signetics (HEF4738) y Motorola 6848. Un interfaz GPIB permite a un ordenador ser talker/listener/controller y ejercer el control del bus GPIB con muy poco software ya que el interfaz maneja todos los protocolos y en general permite interrumpir al procesador activando una línea de interrupción.

El bus IEEE-488 tiene tres características fundamentales que lo hacen especialmente adecuado a un entorno experimental o de laboratorio. La primera es que los equipos, pueden conectarse entre sí con una gran flexibilidad. A cada equipo llega un cable que se conecta mediante un conector tipo Ribbon. Desde este mismo punto, se puede conectar un segundo cable a un segundo equipo, pero como este nuevo cable tendrá nuevamente un macho por una parte y una hembra por la otra, volvemos a tener la posibilidad de seguir conectando equipos. Es decir, de cualquier equipo pueden partir cables a un número variado de otros equipos, desde uno hasta el máximo permitido (15), si aplicamos esto a todos los equipos veremos que la flexibilidad es considerable, aspecto que resulta fundamental en un entorno de laboratorio, donde los equipos se conectan y desconectan de forma frecuente. Otras interfaces, como SCSI, exigen que todos los equipos estén conectados en cadena.

La segunda característica es que los datos que envía un locutor pueden ser recogidos por varios oyentes simultáneamente. Esto permite por ejemplo, que un voltímetro esté enviando datos a un ordenador para análisis y almacenamiento y simultáneamente, esos datos estén siendo recogidos por un filtro digital, o estén controlando la salida de una fuente de alimentación o múltiples cosas de forma simultánea. Para que esto funcione, se emplean las señales de NRFD y NDAC que se conectan al bus mediante la técnica de colector abierto. Según esta técnica, basta que un equipo mantenga su salida en baja, para que la línea correspondiente del bus también lo esté. Esta señal sólo pasará a alta cuando todos los equipos hayan puesto su salida en alta. La señal NRFD indica cuando los equipos están preparados para recibir un dato y la NDAC indica que todos los equipos ya lo han recibido.

La tercera característica, es que permite sincronizar acciones entre distintos equipos, mediante el envío de comandos de disparo. De esta forma, un generador de señal puede comenzar a proporcionar una determinada salida para un equipo bajo estudio y simultáneamente, un

osciloscopio comienza a registrar las señales a la entrada y salida de dicho equipo, y ambas acciones comenzarán en los distintos equipos, y pueden ser todos, de forma sincronizada.

Como se ve son características que lo hacen ideal en entornos de instrumentación y de laboratorio y en este campo, es de utilización prácticamente universal y todos los equipos de laboratorio profesionales lo incorporan o bien de serie o al menos como opción.

Periféricos de entrada

4.1 TECLADOS

Se denomina «teclado» al género de periféricos de entrada, constituidos por un conjunto de botones pulsadores, de tal modo que cada botón se corresponda con un determinado carácter, función, instrucción o idea. El tipo de teclas, así como su número y distribución, vendrán determinados por la aplicación concreta que se desee realizar, por lo que no existen modelos genéricos, sino desarrollos específicos.

El número de teclados conectados a un determinado sistema es sumamente variable, oscilando entre cero (tal como en sistemas muy simples de instrumentación o automatización) y varias decenas (tal como en sistemas multiterminal de recogida y consulta de bancos de datos); no obstante, en sistemas basados en microprocesadores, lo más común es disponer de una unidad, desde donde se suministran al sistema las informaciones básicas en cuanto a selección y control de programas, e introducción de variables.

Físicamente, el teclado acostumbra a ir asociado a otro periférico de salida, tal como una impresora, una pantalla o un visualizador, con lo que el operador obtiene una comunicación bidireccional con el sistema. Este conjunto de teclado y visualizador se trata en muchos aspectos de forma conjunta y habitualmente recibe el nombre de consola.

El teclado, junto con el sistema de vídeo, es el periférico más popular. Es la principal herramienta de entrada de datos al sistema o al menos de su control ya que las grandes cantidades de datos van a través de los sistemas de comunicación o de los dispositivos de almacenamiento intercambiables. Recientemente otros dispositivos como por ejemplo el ratón han ganado popularidad, pero generalmente se usan más para controlar al sistema que para introducir los datos a procesar por el sistema. Necesitamos distinguir entre el teclado estándar, el cual es familiar a todos los usuarios de ordenadores, y varios tipos de teclados especiales, los cuales son diseñados para muchas aplicaciones específicas.

Externamente, el teclado consiste en un conjunto de teclas diseñadas para ser pulsadas por el dedo, y algunas veces uno o más leds luminosos. Internamente el teclado es un circuito electrónico, que detecta cuando cada tecla es pulsada y/o liberada, y envía esta información al procesador principal. A menudo asociamos el teclado con el vídeo, debido a que cada vez que pulsamos una

tecla, aparecerá en la pantalla el carácter correspondiente. Sin embargo, en realidad, esto no ocurre así sino que la señal es transmitida desde el teclado al procesador, y éste entonces escribe el carácter en la pantalla. En el caso de terminales inteligentes, el procesador implicado es comúnmente un procesador dentro del dispositivo. En terminales 'tontos' (carecen de la posibilidad de procesamiento de datos), se depende del ordenador principal, que devolverá el carácter en la pantalla. A veces, este eco puede ser suprimido, como por ejemplo cuando se teclea una contraseña de acceso ('password').

La tecnología empleada en los teclados es relativamente simple. Estos son periféricos que no necesitan de gran sofisticación, porque son utilizados directamente por el hombre y con sus manos, lo que impone restricciones de tamaño, de mecánica, etc. Otra restricción importante es que los teclados comunes de las consolas de ordenador provienen, en su aspecto físico, de la evolución de las máquinas anteriores (máquinas de escribir, teletipos, etc.). Este lastre es evolutivo y muy común en todos los campos de la informática.

El elemento básico de los teclados es el pulsador, elemento electromecánico que conforma la tecla. La idea general consiste en un dispositivo que permanece normalmente en reposo, y que en este estado suministra una señal determinada. Al accionarlo, el movimiento mecánico producido es traducido a una variación de la respuesta eléctrica, que es detectada por alguna circuitería. Este hardware, manejando señales ya digitales, lo comunica al procesador central. A continuación estudiaremos los modelos de pulsadores más interesantes.

4.2 TIPOS DE PULSADORES

El componente básico de un teclado es el pulsador individual. Generalmente, cada tecla controla un simple interruptor que permanece abierto mientras el pulsador está en descanso y se cierra cuando el pulsador es accionado.

Existe un sinnúmero de variaciones, relacionándose más adelante los tipos más usuales. Una primera clasificación entre ellos, se puede hacer atendiendo a la forma en la que se produce el cambio de estado, si es de contacto físico o no lo es. El concepto de conmutación de estado sólido se aplica a veces al describir elementos sin contacto, aunque hablando con propiedad solamente se puede aplicar a una limitada variedad entre la que se cuentan los de efecto Hall o los de elementos fotosensibles.

Dentro de los pulsadores con contacto físico podemos incluir:

- ◆ El de contacto convencional
- ◆ De láminas flexibles
- ◆ De bóveda
- ◆ Reed
- ◆ Elastómeros

y entre los que no tienen contacto físico:

- ◆ Capacitivos
- ◆ Inductivos
- ◆ De efecto Hall

Los pulsadores de contacto son los más simples y baratos y por lo tanto se utilizan bastante. El movimiento que provoca el operario actúa directamente uniendo sus dos contactos que en reposo están separados. Esta unión modifica el nivel eléctrico de uno de los contactos en base al estado del segundo contacto, hecho que es detectado y comunicado al ordenador. El movimiento

mecánico actúa directamente sobre los contactos y permiten, sin grandes complicaciones, configuraciones de contactos múltiples. Los interruptores mecánicos son utilizados básicamente por su economía. El principal inconveniente de los pulsadores de contacto es que éste no es perfectamente instantáneo y suelen producirse rebotes (Fig. 4.1). Un circuito antirebote típico se muestra en la figura (4.2). Este circuito está constituido por dos puertas NAND de dos entradas conectadas formando un biestable tipo RS donde las dos entradas R y S se conectan a estado alto a través de sendas resistencias, lo que garantiza que el biestable mantiene su estado cuando el contacto pasa de una posición a otra. El contacto pone una de las entradas (R o S) en estado bajo y nunca se produce la situación de ambas entradas en estado bajo. Supongamos que inicialmente el pulsador está en la posición A, lo que provoca un estado bajo en una de las entradas de la puerta NAND superior, con lo que independientemente de como esté la otra entrada producirá a la salida (Q) un estado alto. Por el contrario, la puerta inferior tendrá ambas entradas en estado alto y la salida (\bar{Q}) estará en estado bajo. Cuando pulsamos la tecla, el contacto móvil abandona su posición A, con lo que las entradas R y S estarán simultáneamente en estado alto y la salida del biestable no cambia hasta que el contacto móvil llegue a la posición B que producirá un estado alto en (\bar{Q}) y un estado bajo en (Q). Si se produjese un rebote y el contacto móvil abandonase la posición B durante un instante, el estado del biestable quedaría inalterado.

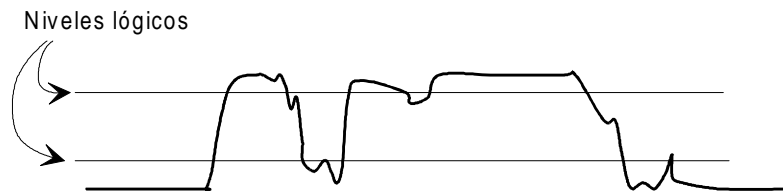


Fig. 4.1 Rebotes en un pulsador de contacto

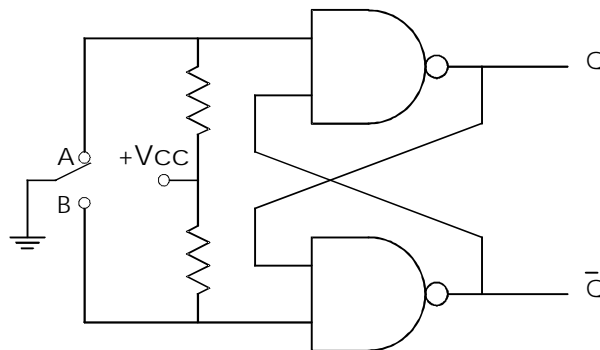


Fig. 4.2 Circuito antirebote típico

Si consideramos a los rebotes como una interferencia o ruido que contamina la señal, podemos eliminarlo con la ayuda de un filtro. En estos casos una sencilla red R-C formada por una resistencia en serie con un condensador en paralelo suele ser suficiente.

Otra forma, también muy corriente de evitar los rebotes, es emplear un monoestable. En este tipo de circuitos el pulsador es utilizado para disparar el monoestable que genera un pulso cuya anchura es independiente del tiempo que permanezca activado el contacto. De esta forma, cuando un pulsador golpea repetidas veces el contacto antes de asentarse definitivamente, la salida es fija, independiente e insensible a estas oscilaciones. Para garantizar esto, la anchura del pulso que proporciona el monoestable debe ser mayor al tiempo de oscilación del contacto, para evitar un redisparo accidental del monoestable.

Otro problema asociado a los pulsadores de contacto es que son sensibles a las condiciones ambientales, como oxidación de los contactos, polvo, humedad, etc, aunque veremos que hay algunos tipos de pulsadores que tienen el contacto en una cavidad sellada herméticamente, lo que obviamente incrementa su coste. Por último señalar que los pulsadores por contacto están sometidos también al desgaste debido al rozamiento y golpeteo de contactos lo que provoca una erosión en el mismo haciéndolo de esta forma muy dependiente del uso que se haga de él. A continuación vamos a ver algunos tipos elementales de pulsadores de contacto y posteriormente veremos los tipos más importantes de pulsadores sin contacto.

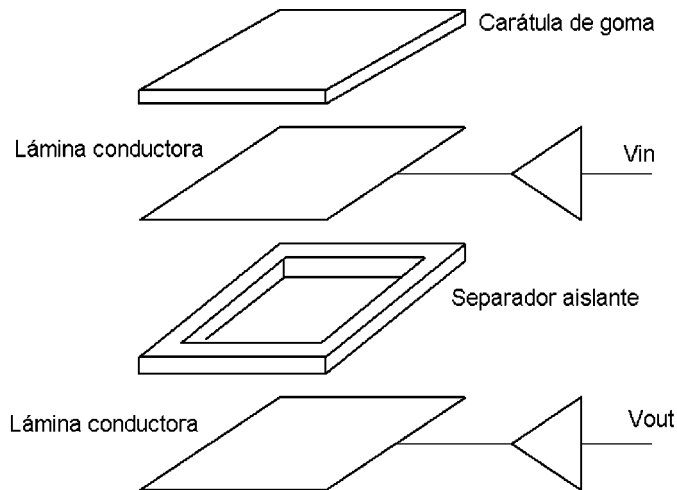


Fig. 4.3 Pulsador de láminas flexibles

4.2.1 Pulsador de lámina flexible

Este tipo de interruptor (fig. 4.3), formado por una serie de láminas sobrepuestas, se basa en la deflexión de un diafragma flexible, metalizado por su cara inferior, que permite establecer contacto con un circuito impreso a través de aperturas practicadas en un separador aislante.

Una cubierta de silicona protege los contactos contra los contaminantes. Algunas versiones más económicas emplean láminas flexibles de silicona conductora, que sustituyen la cubierta protectora y el diafragma metalizado. En algunos casos se emplea una base serigrafiada de tinta conductora como sustrato. No obstante, estas soluciones, aunque resultan mucho más económicas tienen una vida útil mucho más corta debido a la menor resistencia al desgaste de la película conductora frente a la lámina metálica.

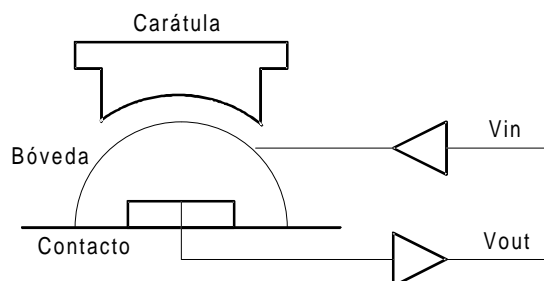


Fig. 4.4 Pulsador de bovedilla

4.2.2 Pulsador de bovedilla

Otro ejemplo de pulsador de contacto es el de bóveda (Fig. 4.4). En él existe una lámina de conductor en forma de bóveda que al ser pulsada, se deforma hasta tocar el conductor que hay

debajo con la cúpula de la bóveda. Al liberarlo, la elasticidad le hace recuperar su forma original y el contacto desaparece. Emiten un clásico chasquido audible que advierte de su correcta operación. Su principal ventaja es que mantienen el contacto sellado y completamente aislado del medio ambiente pero presentan una vida reducida debido a que la metalización no puede ser muy gruesa para permitir la flexibilidad de la bovedilla y se desgasta rápidamente.

4.2.3 Pulsador elastómero

Es muy similar al pulsador de bovedilla. Están constituidos por un circuito impreso, en el que están definidos tantos pares de contactos como debe contener el conjunto del teclado, y un elemento elastómero (silicona) que forma la parte móvil del contacto (fig. 4.5).

La superficie de contacto del circuito impreso está serigrafiada con grafito conductor como protector de la oxidación y mejorador de la conductividad eléctrica, mientras que la parte móvil está formada por una pieza inyectada en material elastómero con un pequeño inserto de silicona conductiva en su centro, de tal modo que al ser pulsada establece conexión entre los contactos definidos en la parte fija del circuito impreso. La sección de la parte móvil es extraordinariamente delicada en cuanto a diseño, puesto que de ella depende la vida útil (número de operaciones) del pulsador y su respuesta táctil.

Este tipo de pulsador ofrece un coste reducidísimo, siempre y cuando el volumen de producción permita anular sensiblemente los costes de amortización de los moldes de inyección, brindando una vida útil comprendida entre 5 a 50 millones de pulsaciones, suficientes para gran parte de las aplicaciones. Es el tipo de pulsador que se emplea habitualmente en los mandos a distancia de los equipos de audio-video domésticos.

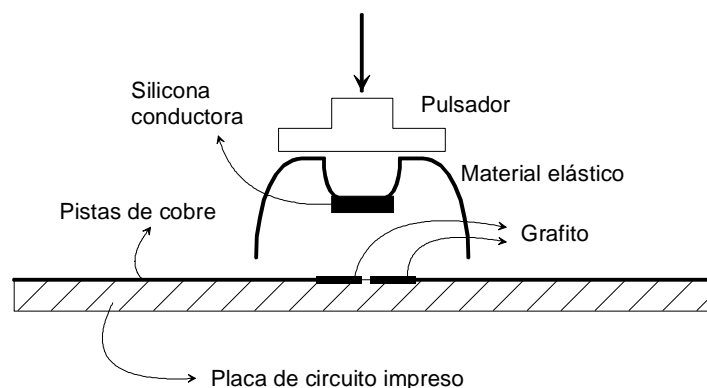


Fig. 4.5 Pulsador basado en elastómero

4.2.4 Pulsadores Reed

El último tipo de pulsador de contacto que veremos es el pulsador REED (fig 4.6). Consiste en una cápsula hermética que alberga dos conductores separados, uno fijo y otro móvil en una atmósfera inerte. Cuando se pulsa, un pequeño imán se aproxima a la cápsula lo suficiente como para atraer al contacto móvil y conectarlo con el fijo. Tiene la ventaja de que el contacto está aislado, volviéndose inmune a la suciedad que es un elemento muy frecuente en todos los teclados especialmente en entornos industriales. Dada la acción indirecta sobre los contactos, no se transmiten sobrecargas mecánicas que provoquen fatiga y desgastes prematuros. Por lo anterior y por la hermeticidad del encapsulado que impide la contaminación de los contactos, este tipo de pulsadores ofrecen una vida útil unas 5 veces superior al clásico pulsador mecánico. El contacto Reed resulta bastante caro y es por lo tanto poco frecuente no siendo rentable fabricar teclados completos, mas bien se emplean para pequeños teclados o contactos aislados en ambientes industriales, donde la duración y fiabilidad son factores predominantes.

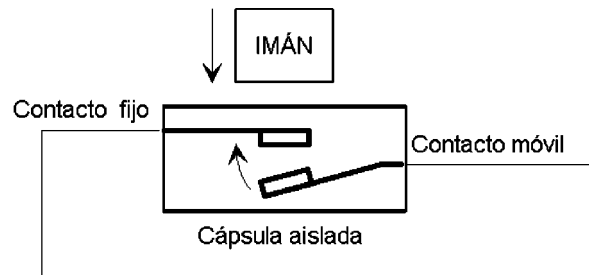


Fig. 4.6 Pulsador Reed

4.2.5 Pulsadores capacitivos

Este tipo de pulsadores emplea un cambio en la capacidad de un condensador para entregar una salida (fig. 4.7). Emplean dos superficies vecinas sobre un mismo circuito impreso, estando una de ellas excitada por la señal alterna de un oscilador; si se aproxima paralelamente una placa conductora sobre ambas superficies, se provoca un acoplamiento entre ellas, con lo que aparece una fracción de la señal alterna de entrada, en la salida.

La señal de salida del pulsador debe ser convenientemente amplificada y convertida a niveles lógicos.

Existen múltiples variantes tales como los que utilizan contactos sensitivos, sin ningún elemento móvil. Otros diseños emplean bovedillas metálicas cóncavas como elemento de acoplamiento. Los pulsadores capacitivos ofrecen la elevada fiabilidad de los interruptores sin contactos móviles.

Dados los bajos niveles de señal entregada por estos pulsadores se presenta una acusada sensibilidad a interferencias y unos serios condicionamientos en la estructura metálica soporte, trazado de pistas en el circuito impreso y electrónica de amplificación, detección y conversión. Por todo ello sólo aparecen disponibles formando parte de teclados completos producidos por fabricantes especializados.

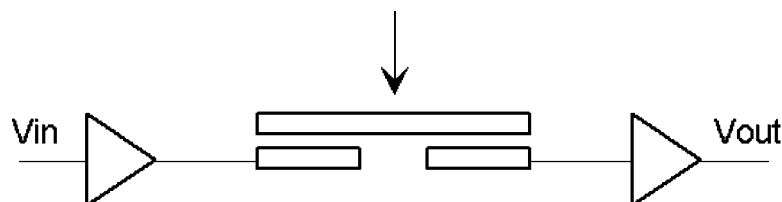


Fig. 4.7 Pulsador de tipo capacitivo

4.2.6 Pulsador de efecto Hall

Los sensores de efecto Hall (fig. 4.8) están formados por un elemento semiconductor recorrido por una corriente continua, y un campo magnético perpendicular a ella que provoca una deformación de las líneas equipotenciales sobre la superficie del semiconductor, apareciendo una tensión de salida proporcional al producto de la corriente de polarización por la intensidad del campo magnético aplicado.

La conmutación se obtiene al aproximar un imán permanente al sensor, que desarrolla una tensión de salida que es amplificada y convertida en digital.

Generalmente el conjunto formado por el sensor, amplificador, disparador de Schmitt, monostable opcional y etapa de salida, forma un circuito integrado monolítico asociado a cada pulsador.

Se distinguen dos tipos fundamentales: estático y dinámico. Los pulsadores estáticos conducen a su salida, mientras exista campo magnético a su entrada. Los pulsadores dinámicos conducen durante cierto período (típicamente 20 ps) cuando el campo de entrada supera el nivel de conmutación, pero no lo hacen durante el resto del tiempo que dicho campo permanezca a nivel elevado, ni durante el alejamiento del imán; para ello incorporan un monostable que dispara en el flanco de subida de la señal magnética.

La salida puede estar formada por un transistor de colector abierto, simple o doble, o bien por una puerta lógica «Y» aceptando señales externas de validación y sincronismo.

Dada la ausencia de contactos, la baja impedancia de todas las señales de interconexión y la insensibilidad a polvo, suciedad y contaminantes, este tipo de pulsadores ofrece la mayor fiabilidad (esencialmente duración infinita), sólo limitada por el desgaste del elemento móvil y el resorte de retorno. Este último, en algunos casos, es sustituido por un sistema magnético de retorno, que proporciona simultáneamente una realimentación al tacto. Su principal inconveniente es su elevado coste.

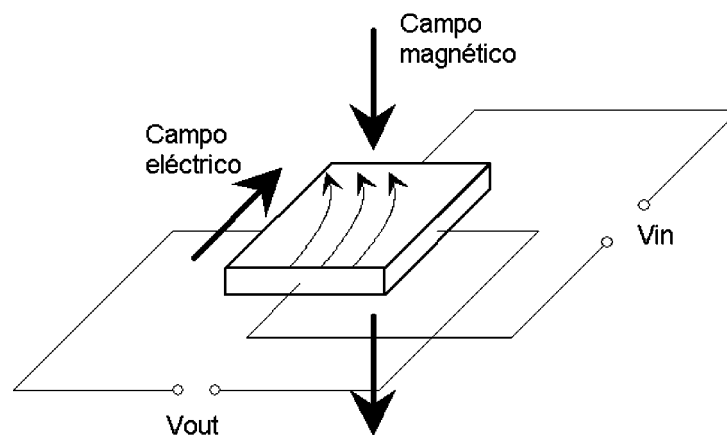


Fig. 4.8 Pulsador de efecto Hall

Basados en el efecto Hall, consisten en la aparición de una diferencia de potencial, cuando un semiconductor es atravesado por una corriente continua en presencia de un campo magnético, según la figura (4.8). De una forma simplificada, podemos describir el efecto Hall de la siguiente forma. Al aplicar una diferencia de potencial entre los extremos del semiconductor, se producirá en su seno una corriente eléctrica, formada por un flujo de electrones y/o huecos. En esta situación, las cargas eléctricas en movimiento se ven afectadas por el campo magnético, que les aplica una fuerza:

$$F = q(\bar{v} \times \bar{B})$$

En donde: q es la carga del electrón
 \times indica un producto vectorial
 \bar{v} es la velocidad del electrón
 \bar{B} es el vector campo magnético

Como el producto vectorial es perpendicular a ambos vectores, se obtiene así una fuerza que desvía a los electrones haciendo que describan una trayectoria curvilínea. La componente

transversal de este movimiento induce en los extremos perpendiculares a la corriente una diferencia de potencial. Esta tensión es amplificada y comparada con una referencia que dará como resultado un estado de 1 ó 0.

El mecanismo actuador del pulsador, lo que hace es acercar dos imanes permanentes al semiconductor hasta que se consigue la tensión de Hall suficiente para ser detectada.

4.2.7 Pulsador inductivo

Este tipo de pulsadores se basa en una variación de la permeabilidad magnética en el medio de acoplamiento de dos circuitos inductivos (fig. 4.9).

Los dos circuitos inductivos están definidos a ambas caras de un mismo circuito impreso. Uno de dichos circuitos está recorrido por una corriente de alta frecuencia (impulsos), mientras que el otro podrá captar dicha señal si existe un buen acoplamiento, y esto se produce cuando la tecla correspondiente es pulsada.

El acoplamiento a través del aire y del propio sustrato del circuito impreso es muy débil, por lo que el nivel de señal obtenido a la salida de los secundarios asociados a las teclas no pulsadas es suficientemente bajo como para ser discriminado como «0» lógico. La pulsación de la tecla introduce un núcleo de alta permeabilidad magnética (ferrita) en el orificio del circuito impreso situado en el eje común a ambas espiras, con lo que el acoplamiento se incrementa de forma notable, induciendo una corriente secundaria (como consecuencia de cada impulso que recorre el circuito primario) de suficiente nivel como para ser discriminada como de nivel lógico alto «1».

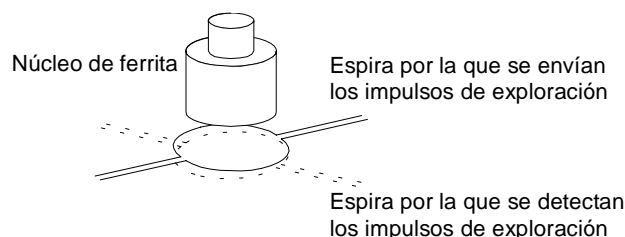


Fig. 4.9 Pulsador de tipo inductivo. Cada una de las espiras, está por una cara de un circuito impreso.

Su fiabilidad es tan elevada como la del resto de técnicas que no implican un contacto físico, es decir: capacitivos, núcleos magnéticos o de efecto Hall. Actualmente se están imponiendo de forma notable dada su combinación de alta fiabilidad y coste reducido, siendo este último consecuencia de no precisar un circuito integrado para cada tecla (como en el caso del efecto Hall), ni un cableado con alta incidencia del coste de mano de obra, ni presentar la sensibilidad al diseño asociada a los pulsadores capacitivos (téngase presente que el circuito correspondiente a cada secundario, usualmente un grupo de 10 a 16 teclas, es un circuito cerrado de muy baja impedancia, por lo que la sensibilidad a interferencias de origen externo está sensiblemente reducida).

4.3 CODIFICACIÓN

Como señales de salida de un teclado pueden utilizarse las conexiones correspondientes a todos y cada uno de los distintos conjuntos tecla-pulsador que lo constituyen. Esto puede ser válido para teclados simples formados por un reducido número de teclas, pero es claramente engorroso si el número total de teclas supera ciertos umbrales.

A título de ejemplo considérese que un teclado mínimo de 12 teclas; no es posible procesarlo mediante una sola palabra en sistemas basados en procesadores de 8 bits (aplíquese lo anterior a teclados alfanuméricos de más de 70 teclas).

Evidentemente se han buscado soluciones mucho más efectivas basadas en la codificación de los datos de salida.

Esta codificación consiste en numerar de forma binaria cada uno de los distintos códigos emitidos por el teclado, de tal modo que el número total de bits precisos para expresar cualquier código no supere los umbrales de maniobrabilidad.

La codificación más usual para teclados numéricos reducidos es la hexadecimal (o su subconjunto BCD si nos limitamos a dígitos decimales); en el caso de teclados alfanuméricos se amplía la codificación, siendo el código más usual el ASCII de 6 o 7 bits según sea reducido o completo, o bien el EBCDIC de 8 bits.

Por circuitos codificadores de un teclado, no solamente se entienden los circuitos precisos para reducir el número de conexiones, sino además el resto de electrónica asociada a teclas y pulsadores; esto incluye, naturalmente, los circuitos destinados a generar las variantes asociadas a cada modo y los destinados a prevenir las pulsaciones simultáneas, que se comentarán más adelante.

4.3.1 Conexión a codificador

Si el número de pulsadores es pequeño, se pueden emplear codificadores para la identificación de la tecla pulsada, como se esquematiza en la figura (4.10) y que consiste simplemente en llevar la señal digital que entregan los pulsadores a un codificador de prioridad BCD que suministra directamente el código de la tecla pulsada. A este circuito se le añaden algunos elementos para generar la validación de la pulsación (VAL). La red formada por la resistencia y el condensador tienen como objetivo retardar la salida de validación de forma que cuando esta se active, los posibles rebotes ya hayan pasado. Con esta sencilla solución, podemos eliminar los rebotes de todas las teclas.

En estado de reposo, cuando ninguna tecla está activa, las entradas del codificador están en estado alto pues están conectadas a la alimentación a través de las resistencias de la parte superior de la figura. Las salidas del codificador están también en estado alto, por lo que la salida de la puerta AND también lo estará. Pasado el tiempo necesario para que el condensador se cargue a través de la resistencia, la entrada correspondiente de la puerta NAND también estará en estado alto, mientras que la otra (B) estará en estado bajo por pasar a través de un inversor. La señal de validación estará en ese momento en alto y el monoestable en su estado estable que podemos suponer, sin pérdida de generalidad, que es el nivel bajo.

Cuando pulsamos una tecla, la correspondiente entrada del codificador se pone en estado bajo y en su salida aparecerá el código correspondiente bajando una o varias de sus salidas. Esto provocará que la puerta AND ponga su salida en estado bajo. Este estado bajo se propaga a la puerta NAND a través de dos canales independientes: por una parte un inversor y por otra a través de la red RC. El cambio en A, aparecerá a la salida del inversor tras un breve instante pasando a estado bajo. Sin embargo la otra entrada requiere que se descargue el condensador y esto llevará algún tiempo que dependerá del producto $R \cdot C$, que se escoge de forma que sea mucho mayor que el tiempo de respuesta del inversor tal y como se muestra en las curvas de la figura (4.10). Esto hace que la otra entrada de la puerta NAND baje lentamente. Durante la descarga del condensador, ambas entradas de la puerta NAND estarán en estado alto y su salida será por tanto un estado bajo. Cuando el condensador se haya descargado casi por completo, la entrada correspondiente de la puerta NAND pasará a estado bajo con lo que su salida conmutará y retornará al estado alto. De esta forma hemos conseguido un pulso que se utiliza para disparar el monoestable.

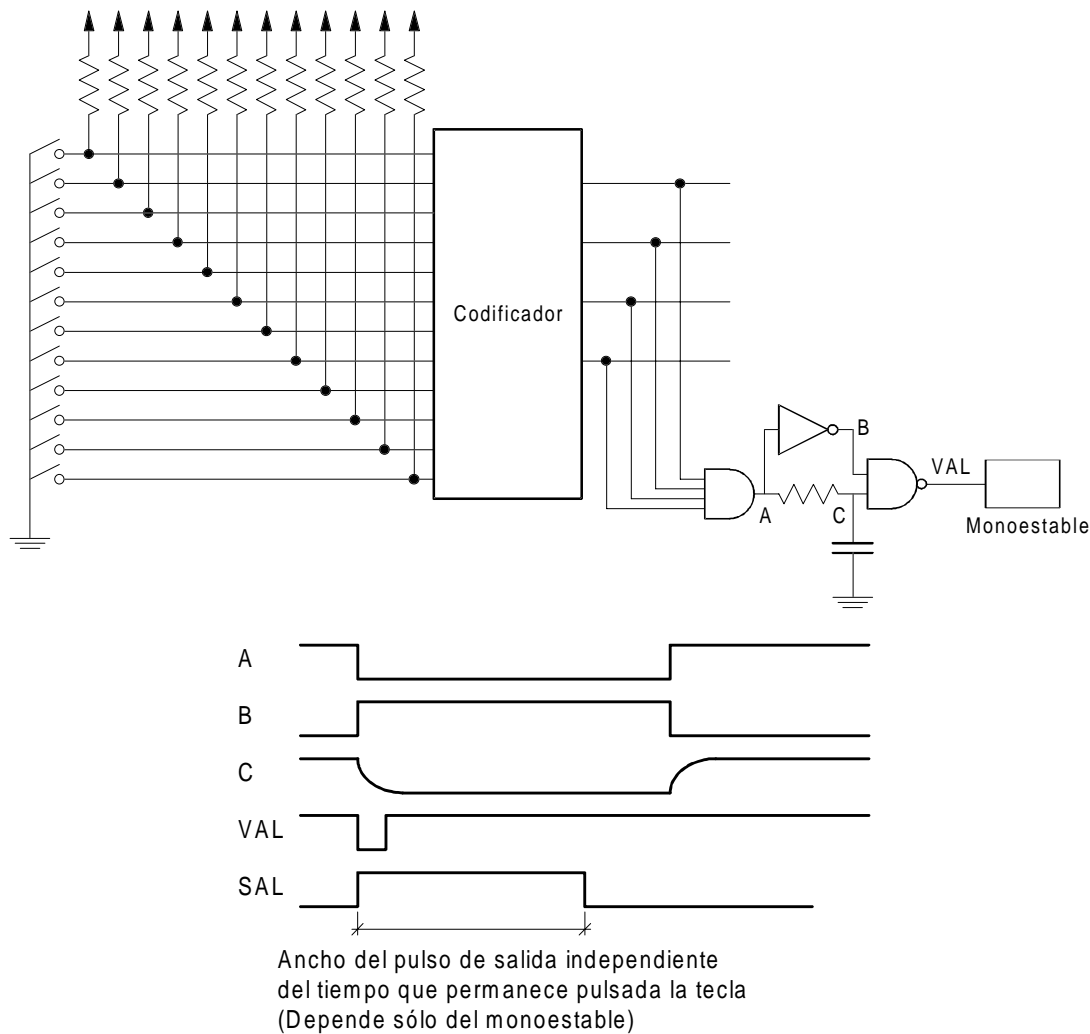


Fig. 4.10 Ejemplo de teclado numérico

4.3.2 Conexión matricial

Cuando el número de teclas sea más elevado (caso típico de teclados alfanuméricos) no es practicable seguir empleando los circuitos codificadores antes descritos, ya que se requeriría un codificador de 80, 100 o más entradas (tantas como teclas). En estos casos el más utilizado, incluso cuando el número de teclas no es elevado, es el teclado matricial (fig. 4.12). El 'truco' está en conectar los pulsadores de forma que el contacto se efectúe entre una fila y una columna de una matriz tal como se refleja en la figura (4.12).

El funcionamiento, es similar al del teclado con codificador que se ha descrito en el apartado anterior. Sin embargo, ahora deben conectarse a tierra dos líneas simultáneamente, para activar los dos decodificadores conectados ahora tanto a las filas como a las columnas. Esto se puede conseguir con un pulsador de doble contacto o con un transistor de doble colector, solución que resulta ideal para los pulsadores de efecto Hall, tal y como se muestra en la figura (4.11).

La técnica más usual consiste en conectar las teclas en forma matricial, de tal modo que el número total de teclas conectable es igual al número de intersecciones. Este tipo de conexión resulta ideal para teclados con pulsadores de tipo inductivo o de efecto Hall dinámico.

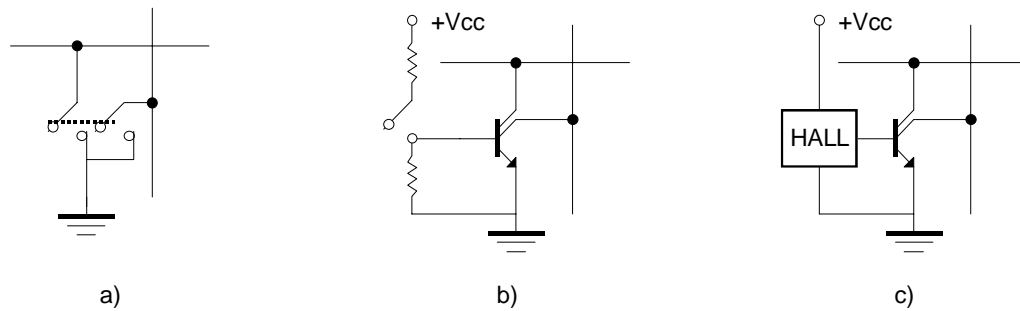


Fig. 4.11 Detalle de una celda de interconexión en un teclado matricial.

a) Con conector de doble contacto, b) con transistor de doble colector, c) con celda de efecto Hall

De esta forma se consigue que el número de terminaciones eléctricas a controlar (filas más columnas) sea sólo del orden del doble de la raíz cuadrada del número de teclas. En el ejemplo de la figura (4.12) se controlan 128 teclas con una matriz de 8 filas por 16 columnas (24 hilos) que suele ser suficiente para la mayoría de las aplicaciones.

Para suministrar al procesador central el código de la tecla pulsada se emplean codificadores de prioridad que generan un código intermedio de 7 bits. Este código se utiliza para direccionar una EPROM en la que están escritos los códigos de las letras. De esta forma, se habilita también la posibilidad de que las teclas cambien el código, como los paginadores, dentro de la EPROM. Añadiendo a este esquema un microcontrolador para comunicar con el procesador central, alguna circuitería adicional y una parte mecánica, obtenemos un teclado muy parecido a los que existen actualmente.

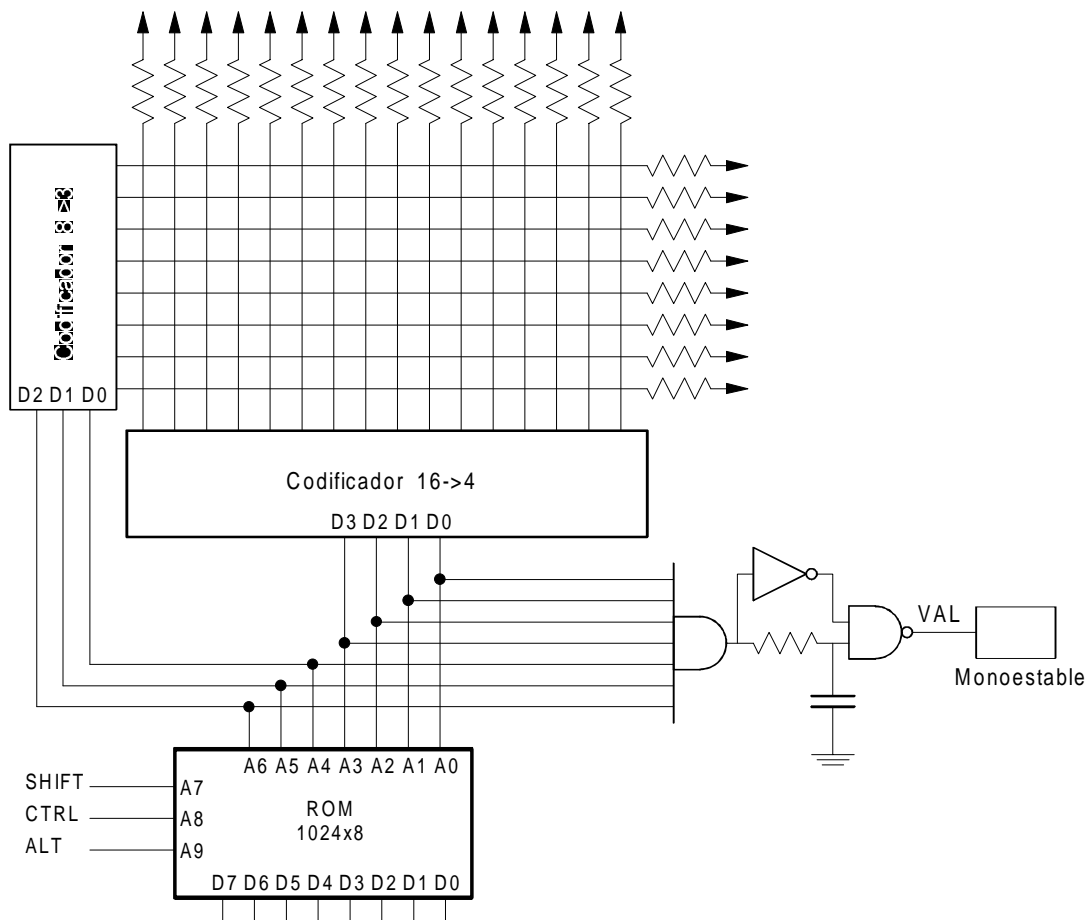


Fig. 4.12 Ejemplo de teclado matricial. Todas las resistencias se conectan a la alimentación (estado alto).

La detección de la tecla pulsada consiste en que al pulsarse una tecla se activan dos líneas a la vez, una correspondiente a una fila y la otra a una columna. Existe un codificador para las filas y otro para las columnas. Las salidas de los decodificadores se conectan a las direcciones de una ROM, junto con las teclas de cambio de modo (SHIFT, CTRL, ALT, etc.), en la que están almacenados los códigos según un determinado sistema (ASCII, EBCDIC, etc.). Problema: la dirección de memoria debe mantenerse cierto tiempo.

4.3.3 Exploración secuencial

En casos donde no resulta conveniente una conexión matricial como sucede con todos los teclados mecánicos así como los Reed se acude como norma general a realizar los circuitos codificadores empleando técnicas de exploración secuencial.

Un circuito clásico es el ilustrado en la figura (4.13) basado en un contador de 7 bits, un multiplexor y un decodificador de 4 a 16. Las teclas codificadas forman una matriz en la que cada tecla conecta una salida del decodificador con una entrada del multiplexor. El decodificador está seleccionado por los 4 bits menos significativos y el multiplexor por los 3 más significativos del contador. Cuando se pulsa una tecla, se cierra una conexión, de tal modo que cuando el contador alcanza el código apropiado, el multiplexor conmuta su salida y dispara un monostable redispensible, que detiene el conteo. El monostable se redispala continuamente mientras la tecla está pulsada con lo que se produce automáticamente la autorepetición.

En el esquema de un teclado con exploración secuencial que se muestra en la figura (4.13), se puede ver que las entradas al multiplexor están en estado bajo debido a que están conectadas permanentemente a tierra a través de las resistencias. De esta forma, la salida del MUX permanecerá en estado bajo sea cual sea la combinación de señales de control. Por otra parte, las salidas del decodificador también estarán en estado bajo excepto la que corresponda al código de entrada que pasará momentáneamente a estado alto durante un ciclo de reloj.

Si no hay ninguna tecla pulsada, las entradas del MUX no se ven afectadas, pero supongamos que se pulsa una tecla. De esta forma conectamos una de las salidas del decodificador con una de las entradas del MUX. La salida del MUX cambiará únicamente cuando estén seleccionadas simultáneamente la salida del decodificador mediante los 4 bits C0-C3 del contador y la entrada del MUX correspondiente a la línea de la columna pulsada. Esta combinación de valores C0-C3, C4-C6 se produce una vez en cada recorrido completo del contador y en ese momento el valor de cuenta del contador nos da directamente el código de la tecla pulsada.

Este código por provenir de un contador binario de 7 bits estará comprendido entre 0000000 y 1111111 pero en la mayoría de las aplicaciones puede interesar otro tipo de código, como por ejemplo el código ASCII con lo que la salida del contador no se utiliza directamente sino que sirve para direccionar una ROM que realiza la conversión de código necesaria. Una PROM o circuitería equivalente realiza la codificación y adaptación de modos, aunque esta función puede ser realizada por el procesador mediante acceso a una tabla residente en memoria RAM y por lo tanto fácilmente modificable por el programa.

Esta memoria ROM que realiza la conversión de código, puede tener además otras entradas adicionales para cambiar el código de salida, tal y como se muestra en las figuras (4.12) y (4.13). De esta forma podemos entender esta memoria como paginada en la que se accede a las distintas páginas mediante esas entradas adicionales y donde cada página contiene un conjunto completo de códigos de tecla.

La conversión que realiza esta memoria ROM, puede hacerla la circuitería integrada en el propio teclado o también, como ya se ha adelantado, un programa y la memoria RAM del sistema, aunque en la mayoría de las aplicaciones relacionadas con ordenadores es una combinación de

ambas. El hecho de que la conversión de código la realice el procesador por software tiene la ventaja de que podemos cambiar la tabla de conversión también por software simplemente cargando un fichero del disco. Esta es la técnica empleada habitualmente para internacionalizar los códigos. De esta forma el teclado devuelve siempre los mismos códigos y el mismo teclado es válido para cualquier país. Para adaptar el teclado a uno u otro idioma basta con cambiar la serigrafía de las teclas y el fichero correspondiente.

Un esquema completo para codificar un teclado de hasta 128 teclas por el método de exploración secuencial resultaría bastante voluminoso si se realiza mediante integrados simples SSI y MSI, máxime si se incorporan circuitos complicados de selección de modos y protectores de sobrepulsaciones. Estos casos han sido resueltos mediante circuitos integrados MOS-LSI, que realizan todas estas funciones.

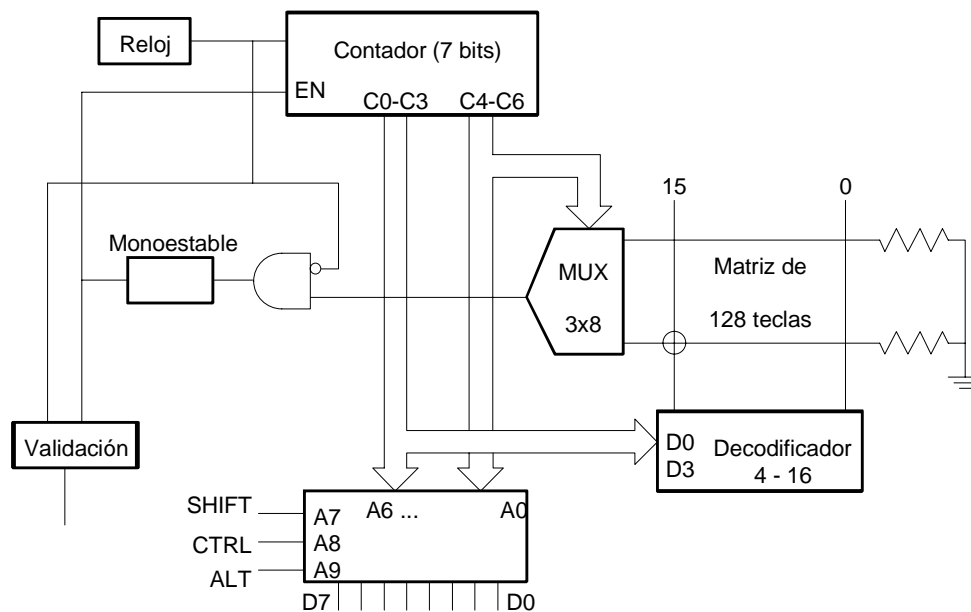


Fig. 4.13 Ejemplo de teclado matricial con exploración secuencial

4.3.4 Codificación por microprocesador

Una técnica que se está expandiendo fuertemente consiste en la utilización de microcontroladores como elementos de lógica activa en la codificación de teclados. Esta técnica permite, con un mínimo de componentes, realizar funciones que hasta el presente raramente eran llevadas a cabo directamente por el teclado como periférico. Entre estas funciones cabe mencionar: exploración secuencial; protección contra pulsaciones simultáneas; transcodificación; modos múltiples; selecciones de modo complejo; salidas en paralelo o en serie; memoria FIFO en caso de pulsación más rápida que el posible acceso por parte de la CPU; autorrepetición en teclas seleccionadas (se denomina autorrepetición al hecho de que si se mantiene pulsada una tecla, pasado un tiempo prudencial, entre 0,5 a 1 segundo, se repiten las validaciones del código asociado a un ritmo aproximado de 10 Hz); autorización o inhibición total o parcial del teclado; generación de señal audible para realimentación acústica; paridad; detección de errores de operación; etc.

Los módulos de pulsador de estado sólido disponen de una entrada y una salida, preparadas para exploración. La señal es válida cuando la entrada de interrogación (exploración) está activada y la tecla pulsada. Dado que el módulo pulsador es un interruptor de estado sólido, con salida digital, podrá conectarse directamente con el microprocesador. No es preciso prever rutinas para eliminación de rebotes, ni circuitos detectores especiales, tales como los que precisan los pulsadores capacitivos o de núcleos de ferrita; con todo ello se dispone de mayor espacio en la ROM para permitir incrementar sus prestaciones.

4.3.5 Doble codificación

El control del teclado puede efectuarse totalmente mediante un único microprocesador, tal como se ha visto en el apartado precedente, o bien distribuyendo el control entre el microprocesador localizado en el propio teclado y el situado en la unidad central. En este último caso pueden emplearse diversas soluciones, siendo la de doble codificación una de las más extendidas.

La técnica de doble codificación consiste en generar dos códigos por cada tecla, el primero en el momento de su pulsación («make») y el segundo en su liberación («release»). Ambos códigos presentan una notable similitud; de hecho sus siete bits de menor peso son habitualmente idénticos, diferenciándose en que el octavo bit es un «0» en la pulsación y un «1» en la liberación o viceversa. Los siete bits de peso inferior acostumbran a codificar de forma binaria la posición física de la tecla, siendo, por tanto, independientes del carácter o función asimilados a dicha tecla; como consecuencia, para obtener el código alfanumérico asociado a la tecla en cuestión, es preciso efectuar una transcodificación mediante el microprocesador de la unidad central. La tarea adicional que comporta esta rutina queda ampliamente compensada por la flexibilidad que se obtiene, puesto que cualquier tecla puede adoptar el significado que se desee, el cual puede ser modificado sin más que alterar la tabla de transcodificación activa. Asimismo, cualquier tecla puede adoptar la característica de tecla de modalidad (tal como: shift, control, alt., etc.) dado que el microprocesador de la unidad central es quien decide, en base a su microprogramación, la funcionalidad asignada a cada tecla. Asimismo, dado que la unidad central de control conoce en todo momento qué teclas están pulsadas (puesto que ya ha recibido el código de pulsación y aún no el de liberación) pueden estructurarse secuencias complejas que comporten combinaciones simultáneas de múltiples teclas, o bien funciones tales como la autorrepetición «typematic» consistente en repetir a una frecuencia de 10 a 15 c/s el código de la tecla cuya pulsación haya sido mantenida un tiempo superior a unos 500-750 ms.

4.4 PULSACIÓN SIMULTÁNEA DE VARIAS TECLAS

Existen diversos métodos para controlar la forma en que el teclado informará al procesador central de cómo y cuando se activan los pulsadores. Las técnicas que se utilizan intentan por un lado evitar conflictos ante el manejo erróneo del teclado, y por otro, agilizar el manejo y evitar esperas superfluas.

El problema se plantea cuando se pulsan varias teclas simultáneamente y, una vez hecho esto, cuando se van liberando sucesivamente.

En los teclados conectados a ordenadores (donde está corriendo un sistema operativo), lo más usual es que el manejador genere una interrupción por cada tecla que se pulse. Incluso, en algunos casos, se genera también una interrupción cuando la tecla se libera, para dotar al manejador de toda la información posible y dejar que actúe como quiera.

Este problema común a cualquier tipo de teclado aparece cuando se pulsan dos o más teclas simultáneamente. Si no se toma ninguna precaución, lo más común es que se provoque una suma inclusiva de bits dando lugar a la generación de un tercer código que no corresponde a ninguna de las dos teclas pulsadas, perdiendo asimismo la información correspondiente a estas últimas.

Aunque pueda parecer que este es un problema de operatoria ajena a los equipos, la tecnología ha desarrollado una serie de soluciones que permiten soslayar los defectos humanos de manipulación. Las soluciones más usuales son las siguientes:

- ♦ sobrepulsación de 2 teclas
- ♦ inhibición de N teclas
- ♦ sobrepulsación de N teclas.

4.4.1 Sobrepulsación de dos teclas

(En inglés: «2-Key rollover».) Cuando se pulsán varias teclas simultáneamente, sólo se transmite el código asociado a la primera, quedando la segunda y consecutivas bloqueadas hasta liberar la primera. Si una segunda tecla fue pulsada tras la primera y liberada antes que ésta, no queda registrada, perdiéndose su información. Este método garantiza que no aparezcan códigos erróneos, pero no impide pérdida de información.

4.4.2 Inhibición de N teclas

(En inglés: «N-Key lockout».) Cuando se pulsán varias teclas simultáneamente no se generan códigos a la salida. Cuando una sola tecla está pulsada, el teclado genera su código, pero cuando se pulsa una segunda tecla mientras la primera permanece activa, el teclado no generará ningún código mientras no se libera la primera. Una vez que se libera la primera, el código correspondiente a la segunda aparecerá a la salida. Por tanto si se pulsán N teclas simultáneamente, permanecerá inhibida la codificación hasta que todas las teclas regresen a la posición de reposo, excepto una.

Este procedimiento es muy similar al de sobrepulsación de dos teclas, diferenciándose por el hecho de que en aquél durante la pulsación múltiple se dispone del código de la primera tecla; mientras que en éste, durante la pulsación múltiple, la salida permanece inhibida.

Merece el mismo comentario respecto a posibles pérdidas de información.

4.4.3 Sobrepulsación de N teclas

(En inglés: «N-Key rollover».) Cuando se pulsa una tecla, se genera su código correspondiente. Si la primera tecla permanece deprimida mientras se pulsa una segunda, se generará la salida correspondiente a la segunda tecla.

Si se pulsa una tercera tecla mientras las dos primeras (o alguna de ellas) están todavía activadas, se genera el código correspondiente a esta tercera tecla.

En un caso extremo, todas las teclas del teclado excepto una pueden ser pulsadas; cuando se activa la última tecla, se generará su código asociado.

Este método se encuentra comúnmente en máquinas eléctricas de escribir, donde ha demostrado su virtud de poder incrementar notablemente la velocidad de tecleo sin generación de errores, ni pérdidas de información.

Generalmente se acepta que los procedimientos de «inhibición de N teclas» o «sobrepulsación de dos teclas» son suficientes cuando aparece una indicación visual, tal como iluminación de una pantalla de TRC, impresión sobre papel o similares. El procedimiento de «sobrepulsación de N teclas», deseable en todos los casos, es absolutamente necesario cuando no se dispone de información visual asociada al teclado.

4.5 RATONES Y TABLETAS GRÁFICAS

Las nuevas tendencias de la programación actual nos hacen trabajar con iconos, ventanas, menús, etc. y se precisa un periférico apuntador: algo que sea capaz de mover gráficamente el puntero que aparece en pantalla. Naturalmente, existen diversas formas de mover ese puntero por la pantalla, que son los periféricos apuntadores: ratones, tabletas gráficas, lápices ópticos, trackballs, pantallas táctiles, etc. Hay claramente dos ganadores: los ratones y las tabletas gráficas de precisión o tabletas digitalizadoras.

Aunque fue Apple quien lo popularizó con su ordenador Macintosh, lo cierto es que la tecnología actual del ratón es obra de Xerox, quienes reinventaron el concepto y los primeros diseños creados por Douglas Engelbert, en el Stanford Research Institute en 1967.

En 1970, Xerox creó el primer ratón digital que incluía circuitos de conversión analógico digitales. El primer ratón para PC fue creado por Mouse Systems, compañía que dio origen a todo un estándar. Microsoft introdujo su primer ratón en 1983, un año antes que Apple en Macintosh.

4.5.1 Funcionamiento básico del ratón

El funcionamiento del ratón es sencillo y ha evolucionado con el paso de los años. Los primeros ratones, electromecánicos dieron paso a los ratones optomecánicos, que son los que se conocen en la actualidad: una combinación mecánica (la bola y los rodillos) con detección digital (fotosensores). Los ratones ópticos son lo último que ha lanzado la tecnología del sector: se utiliza una alfombrilla (o tableta) especial, que contiene una rejilla de líneas dibujadas en ella. El ratón óptico no contiene partes mecánicas, y esa es una gran ventaja puesto que se traduce en menos fallos. A todo esto hay que añadir los botones del ratón: Su número puede variar entre uno (como el ratón Macintosh) y tres. El número normal en los PC's es dos para los compatibles Microsoft, y tres para los compatibles Mouse Systems.

El ratón de tipo optomecánico contiene una bola (fig. 4.14), generalmente de caucho, teflón o goma, que gira soportada sobre tres ejes: Uno horizontal, otro vertical y un tercero oblicuo, que con la ayuda de un muelle sirve para mantener la bola en contacto permanente con los otros dos ejes. Cada uno de los ejes está unido a un pequeño "plato" circular con pequeñas rendijas, que giran en torno a los rodillos al mover la bola (el ratón) sobre la mesa. En la figura (4.14) el tercer eje estaría oculto detrás de la esfera y empujaría la esfera contra los dos ejes que mueven las ruedas ranuradas que interrumpen periódicamente los haces de los fotosensores.

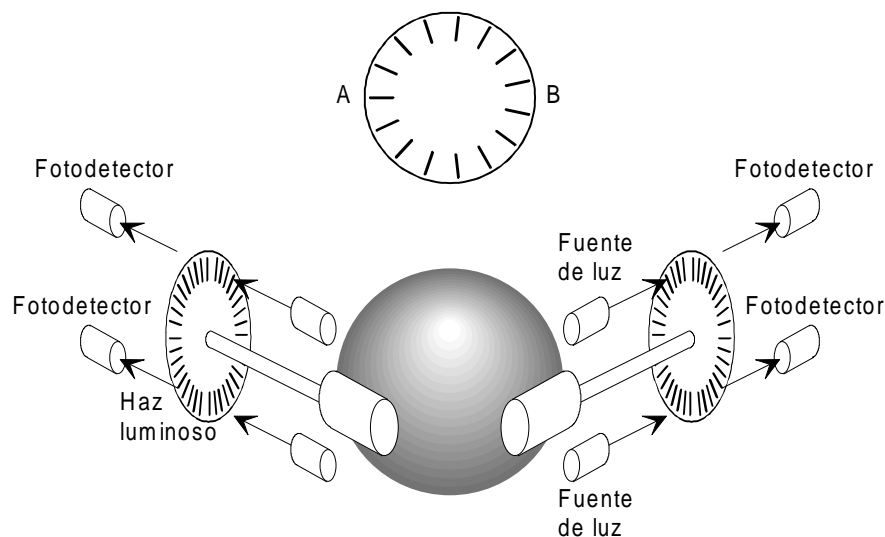


Fig. 4.14 Esquema de un ratón optomecánico

Dos pequeños haces de luz inciden sobre dos fotosensores. La diferencia entre luz-oscuridad en las rendijas debido al giro del plato permite detectar el movimiento, las rejillas están ligeramente desplazadas (en el gráfico puntos A y B); es decir no hay orificios diametralmente opuestos. Esto hace que uno de los dos fotosensores se dispare el primero indicando el sentido de giro, como se muestra en la figura (4.15). La información se convierte en bits en función del movimiento y se transmite al ordenador para desplazar el puntero por la pantalla.

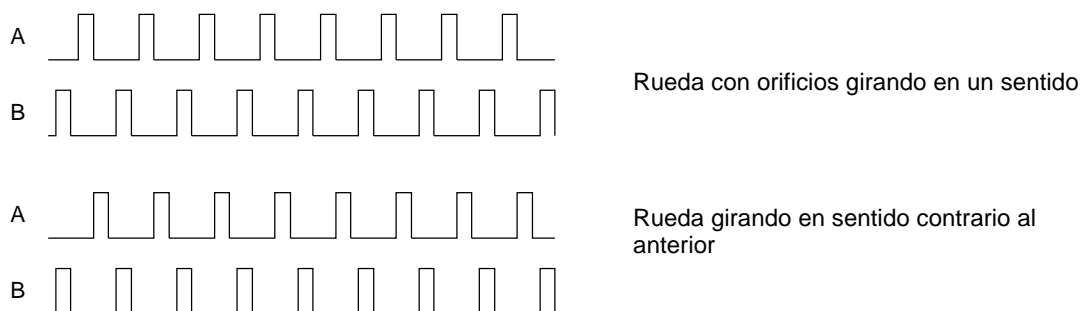


Fig. 4.15 Forma de los trenes de pulsos de los fotodetectores que permiten averiguar el sentido de giro.

La tecnología de los ratones ópticos es completamente diferente. La parte interior del ratón contiene dos agujeros y dos focos de luz (A y B) que emiten sendos haces, que reflejados en la alfombrilla especial, y pasando a través de unas lentes, son detectados por un par de fotosensores (C y D). Los fotosensores pueden medir los pasos de luz (oscuridad), reflejados entre las líneas o puntos de las rejillas, convirtiéndolos en información sobre el movimiento. Hay distintas tecnologías basadas en este sistema: Unas utilizan luz polarizada y otras luz roja e infrarroja sobre una alfombrilla con líneas negras y azules, para distinguir desplazamientos horizontales y verticales.

4.5.2 Nuevos ratones

En la actualidad están surgiendo nuevos tipos de ratones que aún no tienen mucha implantación en el mercado. Un modelo de ratón curioso es el 'ratón a distancia', el cual es un ratón convencional que en vez de cable utiliza infrarrojos para transmitir las señales al ordenador. Este tipo, evita el molesto cable, pero su campo de acción sobre la mesa está limitado, debido a que necesita una línea recta libre de cualquier obstáculo entre él y el receptor que está conectado físicamente al ordenador ya que al permanecer estático no es preciso que sea un elemento independiente con lo que se elimina un dispositivo y la molesta conexión al sistema central. El principio de funcionamiento es el mismo que en los mandos a distancia de los equipos de audio y vídeo doméstico.

Más recientemente se han introducido los ratones inalámbricos que emplean una señal de radiofrecuencia, con lo que no se necesita una línea recta despejada entre el ratón y el receptor. Estos ratones pueden actuar desde cualquier posición de una habitación y esto supone un problema cuando varios ratones de este tipo, conectados a distintos equipos deben convivir en un entorno cercano. Para solventar este problema emplean múltiples canales de emisión de forma que cada par emisor-receptor disponga de un canal exclusivo. De esta forma no se producen interferencias.

Otro modelo de ratón es el 'ratón lápiz'. Es un mini-ratón con una minibola colocada en la punta de un lápiz grande. Se puede manejar como un lápiz sobre la mesa o los papeles, y la bola rueda de la misma forma.

Los ratones denominados "trackball" son una extraña variante de los ratones. Si cogemos un ratón, le damos la vuelta y movemos la bolita con el dedo, ya tenemos un trackball. Son tan similares que la mecánica interna es prácticamente la misma. El trackball puede instalarse sobre una consola de ordenador, de modo que no es necesario el cable que conecta el ratón a la máquina, por lo que se ha popularizado entre los ordenadores portátiles.

En este tipo de equipos se han introducido recientemente dos tipos de ratones de reducidas dimensiones y ausentes de partes mecánicas. El modelo popularizado por Toshiba e incluido en todos sus portátiles, incorpora entre las teclas un pequeño apéndice montado sobre unos pequeños elementos piezoeléctricos. Los elementos piezoeléctricos proporcionan una tensión eléctrica cuando son deformados en una determinada dirección. Si colocamos dos elementos con sus

direcciones más sensibles en sentido perpendicular, ya tenemos una forma de determinar la dirección y sentido del desplazamiento.

El otro modelo "touch-pad", consiste en un pequeño rectángulo sensible al tacto. Al mover el dedo sobre este rectángulo el dispositivo lo detecta y envía al sistema la información correspondiente. Al igual que con las pantallas táctiles donde podemos encontrar sus antecesores más directos, son muy diversas las tecnologías que se emplean para detectar la posición del dedo.

4.5.3 Tabletas gráficas

Desde el inicio de la informática actual, las tabletas gráficas, -también llamadas tabletas digitalizadoras o simplemente digitalizador- han tenido gran importancia en muchos campos de aplicación, pero especialmente en dibujo y diseño asistido por ordenador (CAD).

El trabajo realizado por estas tabletas, consiste en transformar en imágenes los dibujos que se realizan sobre ellas y presentarlas en la pantalla.

Las tecnologías de las tabletas gráficas, al igual que los ratones han evolucionado con el paso del tiempo. En las primeras tabletas sensoras, el usuario debía presionar el lápiz sobre la superficie, lo que producía el consiguiente desgaste de la tableta; además la precisión, o número de puntos detectables, no era muy grande. Actualmente, las tabletas pueden utilizarse con un lápiz normal (presión), con un ratón especial (campos magnéticos) o con un lápiz propio (señales transmitidas por cable). En muchos casos, la precisión obtenida es del orden de 1000 puntos por pulgada. Incluso se puede dibujar sin tocar la tableta con el lápiz: un papel colocado encima no parece importarles a las tabletas actuales, lo que puede ser de gran utilidad en muchos casos.

Las últimas tabletas incluyen un lápiz o ratón sin cable, que proporciona una apariencia más real a la hora de dibujar en ellas. En muchas ocasiones las tabletas pueden emular el funcionamiento de los ratones por software para garantizar su compatibilidad.

La ventaja de las tabletas es que son periféricos de movimiento absoluto, frente a los ratones que son de movimiento relativo. Un simple 'click' en cualquier lugar de la tableta y el puntero saltará hasta allí automáticamente, independientemente de su posición inicial.

Funcionan detectando la posición absoluta del cursor o lápiz sobre su superficie. Los modelos antiguos detectan la presión de un lápiz sobre una rejilla de diminutos contactos situada bajo la tarjeta. Esto permite calcular rápidamente las coordenadas x e y sobre las que se encuentra. En los modelos más actuales, el lápiz o cursor emite señales o crea un campo magnético sobre la tableta, con lo cual también se puede detectar su posición. Estos sistemas permiten escribir sin ni siquiera tocar la tableta. Este hecho ofrece la posibilidad de interponer un documento entre la tableta y el lápiz.

4.6 LECTORES DE CÓDIGO DE BARRAS

El código de barras es una de las técnicas de recogida de datos usada con ordenadores de más rápido crecimiento aunque en sectores bastante específicos. Existen escáner ópticos que han llegado a ser muy familiares debido a su uso en supermercados y en otros pequeños comercios. Tienen también un gran campo de aplicación en la industria, transportes de mercancías, sistemas automatizados de almacenaje, etc. Se han diseñado de modo que deben recoger la información que se halla impresa como una secuencia de barras de ancho y espaciado variable, u ocasionalmente como una serie de círculos concéntricos. Un código de barras simple se muestra en la figura (4.16). El código de barras puede verse como una versión ampliada del código Morse con barras estrechas representando los puntos y barras anchas representando las rayas. No obstante, la analogía no es del todo precisa ya que en el caso de los códigos de barras, los huecos o espacios en

blanco también se utilizan para almacenar información, y además, tanto las barras como los espacios pueden ser de distintos anchos y no sólo anchos y estrechos. Las barras se leen haciendo un único registro de una línea (figura 4.16). El medio en el que se encuentra el código de barras es mucho más variable que en muchas otras aplicaciones de barrido y en particular, no es necesario que sea plano. Los códigos de barras pueden usarse en latas, botellas y bolsas de plástico, y por supuesto, también en papel. Sin embargo, la especificación de los propios códigos de barras se define estrictamente y deben ser impresos con cierta precisión, algunos tipos antiguos de impresoras por ejemplo, no tienen la suficiente definición. Existen varios conjuntos de códigos (incluyendo desde los que se usan en almacenes hasta los que se usan para libros u otras aplicaciones específicas), pero en general lo principal es codificar cada producto de manera única. La codificación se realiza usualmente utilizando los primeros caracteres de código para identificar un fabricante y éste asigna los restantes caracteres según considera adecuado. Esto implica que el código de barras en sí mismo no transporta información significativa (aunque existen excepciones) sino que se usa normalmente de forma indirecta para direccionar información almacenada en una tabla.

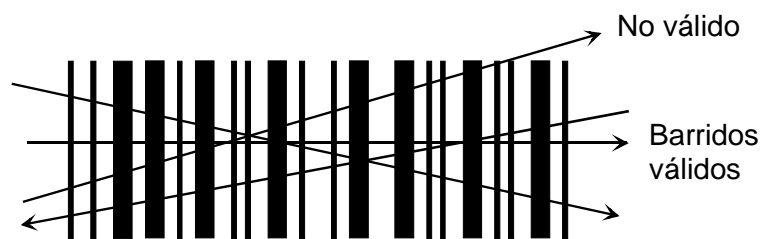


Fig. 4.16 Ejemplo de código de barras simples, sobre el que se muestran varias trayectorias de barrido válidas y una no válida debido a que no recorre todas las barras.

El código de barras es una tecnología de identificación automática que permite recoger datos en tiempo real de manera exacta y rápida pero por sí mismo no resuelve problemas. La combinación de códigos de barras con el hardware y las aplicaciones de software apropiados permitirá la mejora del funcionamiento, de la producción y en último término, de la rentabilidad.

Los símbolos de código de barras pueden imprimirse a bajo coste usando una amplia variedad de técnicas de impresión, y todo el símbolo puede ser escalado uniformemente aumentando o disminuyendo su tamaño según los requisitos de cada aplicación. El código de barras es una tecnología unidimensional (sólo el ancho de las barras y los espacios contienen información). La altura de estos elementos puede considerarse como una medida de la redundancia de datos de los símbolos de código de barras. No obstante, existen códigos de barras bidimensionales e incluso basados en círculos concéntricos aunque en este último caso la información sigue siendo bidimensional.

Los sistemas de código de barras ofrecen a menudo una seguridad muy alta de los datos, el error de sustitución puede ser a menudo mejor que 1 error en un millón de caracteres.

4.6.1 Simbología de códigos de barras

Simbología es el término usado para describir reglas no ambiguas que especifican la forma en que se codifican los datos según el ancho de barras y espacios. Esta codificación se realiza de forma análoga al lenguaje. Cuando nos comunicamos por medio del habla o la escritura podemos usar cualquier lengua siempre que ambas partes estén de acuerdo sobre el idioma elegido y lo conozcan. Lo mismo sucede en el uso del código de barras y dependiendo de los datos que vayan a comunicarse se usarán diferentes simbologías. La comunicación obviamente no se producirá a menos que los equipos de lectura e impresión usen una simbología compatible. Esta consideración

es importante ya que existen docenas de simbologías de código de barras diferentes desde el comienzo de esta tecnología.

A continuación veremos los parámetros y características de la simbología de códigos.

- ◆ Conjunto de caracteres. Este término describe el rango de caracteres de datos que puede codificarse dentro de una simbología dada. Algunas simbologías pueden codificar sólo números denominándose por tanto simbologías numéricas. Otras simbologías pueden codificar información alfanumérica mientras que otras soportan códigos únicos de 128 enteros o el conjunto de caracteres ASCII.
- ◆ Tipo de simbología. La simbología del código de barras puede dividirse en dos categorías generales: discreta y continua. En el primer caso cada carácter se coloca sólo y está separado de caracteres vecinos mediante un hueco intercaracteres. El ancho del hueco no contiene información. En el proceso de decodificar cada carácter es tratado individualmente. En el segundo caso no existen huecos intercaracteres sino que cada carácter comienza con una barra y termina con un espacio. Debido que no existen huecos intercaracteres, un símbolo continuo requiere menos longitud para codificar una cantidad dada de datos. Contrarrestando estas ventajas en la densidad, está el hecho de que el rango de tecnologías de impresión demandadas disponibles es algo más restringido para códigos continuos que para simbología discreta.
- ◆ Anchura del elemento. En un símbolo de código de barras el dato es almacenado en los anchos de las barras y espacios. Existen dos tipos básicos de código de barras: los que emplean sólo dos anchos de elementos (estrecho y ancho) y los que usan anchos múltiples.
- ◆ Longitud variable o fija. Algunas simbologías por su estructura misma codifican sólo mensajes de longitud fija. Otras simbologías se usarán en entornos de longitud fija debido a consideraciones de seguridad de los datos mientras que otras podrán usarse para codificar datos de longitud variable.
- ◆ 'X' y 'Z'. 'X' es el término usado para describir el ancho nominal de los elementos estrechos de un código de barras (tanto barras como espacios). Cuando se examina un símbolo desconocido, es bastante común medir y calcular el ancho medio de los elementos estrechos, esto estrictamente hablando no es 'X' sino 'Z'.
- ◆ Densidad. Las simbologías de códigos de barras se diferencian en la cantidad de datos que pueden codificarse en una longitud dada. Para poder hacer comparaciones significativas debe tenerse en cuenta el valor de 'X' cuando se examinan densidades relativas. Hemos de señalar que la densidad está normalmente especificada sólo por los caracteres de datos. La longitud completa del símbolo incluye caracteres de comienzo/parada, zonas vacías y caracteres de chequeo.
- ◆ 'Self-checking'. Una simbología se denomina 'self-checking' si un único defecto de impresión no deja que un carácter sea transpuesto en otro carácter válido en la misma simbología.
- ◆ Código de comienzo, código de parada. Un código de comienzo es un patrón particular de barras y espacios que está situado al comienzo de un símbolo para indicar al escáner dónde comienza el código. El código de parada se encuentra al final del código para indicar el final de los caracteres de datos. Permiten además detectar fácilmente el sentido de lectura (directo o inverso)
- ◆ Carácter de chequeo. Un carácter de chequeo es un carácter (o caracteres) situado en una posición predeterminada en un código y cuyo valor está basado en algunas relaciones matemáticas de los otros caracteres del código. El escáner usa este carácter para validar que el dato correcto ha sido decodificado.
- ◆ Bidireccional. Una simbología es bidireccional si el símbolo puede barrerse indistintamente a izquierda o derecha sin que ello afecte a los datos decodificados. Casi todas las simbologías que se usan actualmente son bidireccionales. Para esto se requiere, o que haya códigos de inicio y final o que no haya caracteres que tengan una representación simétrica de barras y espacios.
- ◆ 'Self-clocking'. (autoreloj) Los escáners necesitan una información de referencia con objeto de tener una forma de medir la posición relativa de los bordes de todos los elementos. Algunas viejas simbologías incluían una pista de reloj separada. Las simbologías modernas están

diseñadas de tal forma que el escáner no necesita una pista de reloj separada para recobrar la información sobre el ancho (propiedad de 'self-clocking').

Algunas simbologías que se usan en la actualidad son la UPC ('Universal Product Code') que usa un código de 10 dígitos y se usa para identificar únicamente un producto y su fabricante. Otras simbologías estándar de la industria son 'Interleaved 2 of 5', 'Codabar', 'Code 128', 'Code 93', 'Code 49', 'Code 39', PostNet o 'Code 16K'. También hay códigos bidimensionales y circulares.

4.6.2 Equipamiento de lectura

Un lector de código de barras es un dispositivo que se usa para extraer la información que está codificada ópticamente y la convierte en datos digitales compatibles con el ordenador. El lector de código de barras necesita realizar siete funciones básicas para decodificar la información de un símbolo de código de barras:

- 1.- Encontrar los elementos correctos.
- 2.- Determinar los anchos de cada una de las barras y espacios del símbolo.
- 3.- Cuantificar los anchos de los elementos en un número de niveles apropiado a la simbología que se esté usando (dos para el 'código 39', 'interleaved 2 of 5', cuatro para 'UPC' y 'código 93', cinco para 'código 128', etc.).
- 4.- Asegurar que los anchos de los elementos cuantificados son consistentes con los de las reglas de codificación de la simbología usada. Comparar el patrón de los anchos de los elementos cuantificados con una tabla de valores almacenados para esa simbología y determinar los datos codificados.
- 5.- Si es necesario, cambiar el orden de los datos. La dirección leída se determina examinando los caracteres de comienzo/parada.
- 6.- Confirmar que existen zonas vacías válidas y los dos extremos del símbolo.
- 7.- Confirmar que cualquier carácter de chequeo es consistente con el dato decodificado.

El segundo paso, es decir, la medida del ancho de los elementos se realiza por medio de un sistema de barrido electro-óptico en combinación con el software del microprocesador del lector de código de barras. Los siguientes cinco pasos se realizan por software implementado para un algoritmo particular de decodificación.

Un lector de código de barras puede considerarse como dos elementos separados: el dispositivo de entrada y el decodificador. Estos dos elementos pueden estar separados y conectados por un cable o pueden estar en una única unidad.

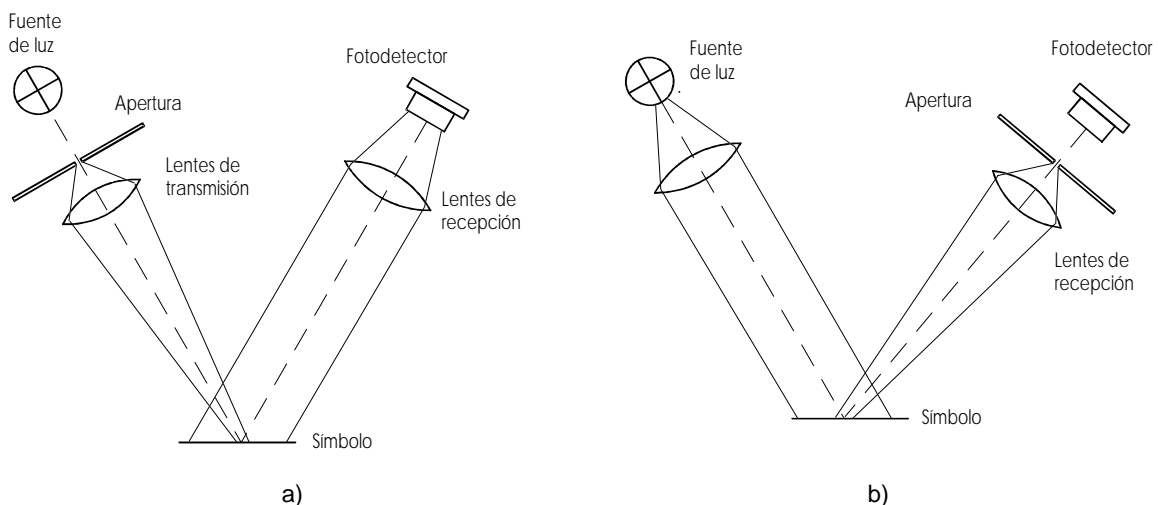


Fig. 4.17 Focalización en emisión (a) y en recepción (b)

El dispositivo de entrada emplea técnicas electro-ópticas para barrer el código de barras. El dispositivo ilumina el símbolo con energía luminosa y examina la cantidad de luz reflejada por un área localizada del símbolo. Normalmente se emplea un laser de semiconductor en los más económicos o los de tipo manual, o de gas en los estáticos y de elevadas prestaciones. Nos referiremos a esta pequeña zona o área como punto y no tiene que ser necesariamente circular. Las dimensiones del punto en un eje perpendicular al eje de la longitud de las barras debe ser consistente con el ancho del elemento más estrecho que sea barrido. El punto se forma bien a partir de un amplio conjunto de luz reflejada desde un haz focalizado (figura 4.17 a) o iluminando el símbolo con luz y usando una apertura focalizada para recoger la luz (figura (4.17 b). Por supuesto se puede focalizar en emisión y recepción, pero en este caso se necesitaría un gran alineamiento y el código siempre tendría que colocarse en la focal de los dos sistemas para poder ser leído.

La luz reflejada del punto del símbolo es dirigida a un detector (fotodiodo o algún dispositivo equivalente) que genera una pequeña corriente proporcional a la luz recibida. Un amplificador en la entrada del dispositivo amplifica la señal desde el fotodiodo hasta un nivel adecuado de modo que pueda utilizarse. El voltaje analógico recogido por el detector es proporcional a la reflectividad observada por el punto de barrido ya que el haz de barrido cruza el patrón del código de barras. Nótese como la forma de los bordes del patrón del código de barras en la señal a la salida del detector es redondeada (figura 4.18.a). Esto es consecuencia del ancho finito del punto que hace que al pasar de una barra (oscura) a un espacio (claro) el punto no lo hace de forma instantánea sino que habrá un espacio de tiempo durante el cual, el punto luminoso incide sobre la barra y sobre el hueco simultáneamente. Por este motivo se debe conseguir un haz estrecho; cuanto más estrecho sea el punto más abrupta será la transición. Con objeto de diferenciar entre barras y espacios la señal analógica es convertida en una señal digital por medio de un circuito conocido como 'wavershaper'. Esta operación se ilustra en la figura 4.18.b.

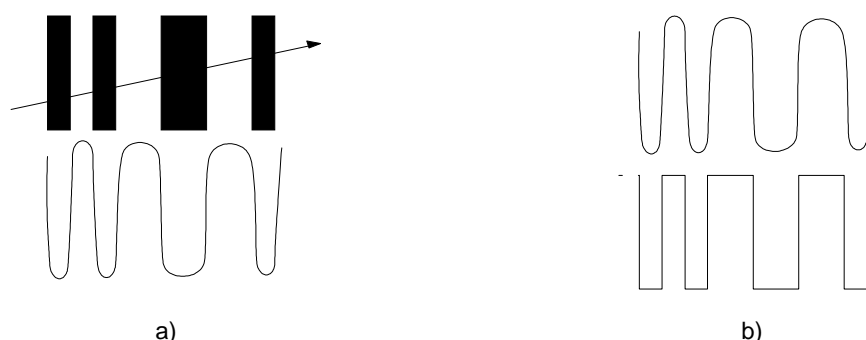


Fig. 4.18 a) Barrido de un símbolo (y representación simplificada de la salida del detector)
b) Operación de regeneración de la señal para obtener una secuencia digital (0 y 1)

El voltaje de salida de un dispositivo de entrada puede ser analógica o digital. Si la salida es analógica debe incorporarse un circuito 'wavershaper' en la unidad del decodificador. La figura (4.19) muestra el diagrama de bloques completo del dispositivo de entrada.

Existen muchos tipos de lectores de códigos de barras. El primero, descrito como un lápiz óptico activo ('active light pen' o 'wand'), está hecho con forma de un lápiz y tiene un único fotodetector junto con una fuente de luz LED en su punta. Ésta está conectada a través de un cable a la parte estática del dispositivo. El operador pasa la punta a través del código de barras y el dispositivo detecta las barras claras y oscuras y las traduce en caracteres. El código incluye también un carácter de chequeo para comprobar que el código se ha leído correctamente y se lo indica al operador usualmente con un tono audible. No se necesita mover el lápiz a velocidad constante y siempre que el movimiento no sea demasiado irregular, puede reconocerse el patrón de barras. Además, el código puede barrerse en cualquier dirección, siempre que se recorran todas las barras (fig. 4.16).

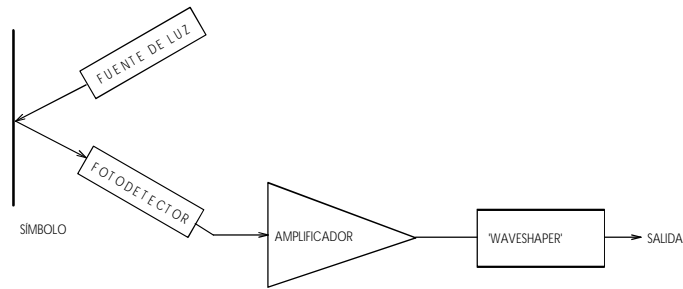


Fig. 4.19 Diagrama de bloques del dispositivo de entrada

El segundo tipo de lector es también bastante común. En este caso, el rayo del LED (o un láser semiconductor) está diseñado para barrer repetidamente a lo largo de una línea en lugar de ser un único punto. El operador sólo tiene que situar el escáner enfrente del código de barras y aproximadamente paralelo a éste y no necesariamente en contacto. El barrido del haz se consigue con la vibración de un espejo que refleja el haz.

El tercer tipo de lector de código de barras es estacionario, usualmente está construido dentro de una caja de chequeo y debajo de un panel de cristal. De nuevo se barre el haz pero esta vez sigue un curso complejo trazando líneas en cuatro direcciones usualmente, situadas entre ellas a 45 grados para lo que se emplea un prisma giratorio. El código de barras no necesita estar en una superficie paralela a la parte superior de la caja ya que siempre que pueda 'verse' por el escáner, puede leerse. Si varios códigos de barras están impresos en el paquete, de modo que siempre hay uno que puede verse por el escáner, el código puede pasarse sobre el escáner en cualquier orientación, aunque esto no es frecuente. De nuevo, el dispositivo indica al operador si el código se ha leído correctamente. Este es el tipo de lector que podemos encontrar habitualmente en los supermercados.

Generación de vídeo

5.1 INTRODUCCIÓN

El vídeo es uno de los periféricos del ordenador más comunes y ciertamente junto con el teclado al que más atención por término medio le muestra el usuario.

La salida visual en los primeros computadores era simplemente una serie de indicadores luminosos individuales. Si el computador y el usuario necesitaban comunicarse caracteres se usaba una impresora y un teclado. Los indicadores luminosos fueron más tarde sustituidos por dispositivos que podían mostrar caracteres o filas de ellos. Estos dispositivos se siguen usando en calculadoras, relojes digitales, paneles publicitarios, etc. Pero en los ordenadores evolucionaron a pantallas, que pueden mostrar caracteres en filas y columnas y que además pueden ser gráficos, aprovechando el desarrollo alcanzado por estos dispositivos gracias al auge de la televisión.

El más común de los tipos de pantallas está basado en el tubo de rayos catódicos (CRT), el cual es usado también en la televisión. Este es todavía el método más barato para producir unas imágenes de calidad, bien sean de color o monocromo. Sin embargo el gran tamaño de éstos es una desventaja.

La relevancia de los monitores de TRC está avalada, por diversos motivos:

- ◆ Madurez en la tecnología (la misma que la televisión).
- ◆ Característica interactiva.
- ◆ Posibilidades de color y realismo, sin comparación hasta el momento.
- ◆ Buena relación coste/prestaciones.
- ◆ Silencioso.
- ◆ No requiere consumibles.
- ◆ Se adapta fácilmente a distintos modos de presentación.
- ◆ Enlaza directamente con las aplicaciones multimedia cada vez más difundidas.

Tras numerosos intentos infructuosos de hacer plana la pantalla del televisor, hay pantallas planas en el mercado y se utilizan de forma habitual en ordenadores portátiles donde el diseño compacto es vital, pero son caros y ofrecen un resultado menos satisfactorio que los TRC.

5.2 GENERACIÓN DE LA IMAGEN EN UN TRC

Como se ha comentado, el más popular de los mecanismos para generar la imagen es el TRC (Tubo de Rayos Catódicos) que se muestra de forma esquemática en la figura (5.1) y que tiene un principio de funcionamiento idéntico al del tubo de los televisores convencionales. Tiene una gruesa pantalla de cristal, casi plana, delante de un armazón en forma de embudo, que se extiende hacia atrás hasta un estrecho cuello de botella. Los tubos pueden ser de varios tamaños, pero el más popular es el de 14 pulgadas aunque recientemente, en aplicaciones profesionales a quedado desbancado por los monitores de 17 pulgadas. Esta medida se toma a lo largo de la diagonal de la pantalla al igual que sucede en los televisores. Para conocer las dimensiones ancho y alto basta con tener en cuenta que la relación entre ellas es $4/3$. De esta forma, conociendo la hipotenusa de un triángulo rectángulo, y la relación entre los lados, se puede obtener con un sencillo cálculo de geometría elemental, las dimensiones laterales. Para un monitor de 14" el área de vídeo útil es de unas 10" de ancho y 8" de alto.

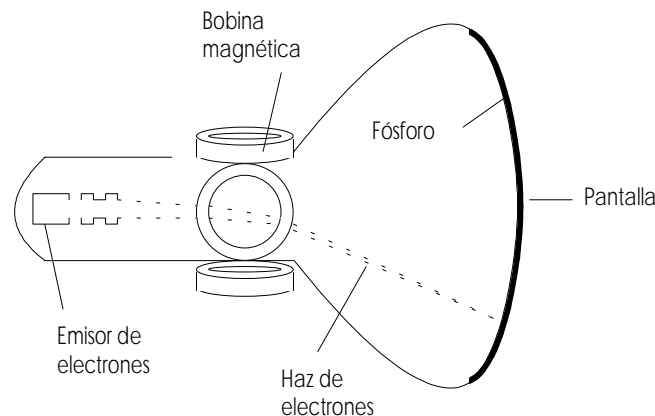


Fig. 5.1 Tubo de rayos catódicos típico

El rasgo esencial de los TRC es un haz de electrones o rayo catódico (porque es emitido por el cátodo). Este haz es emitido por un cañón de electrones en el estrecho cuello al final del tubo, y pasa a lo largo del armazón del tubo, hasta golpear el frontal de la pantalla.

El cañón de electrones consiste en un número de placas de metal, o electrodos, con un cuidadoso control de voltaje aplicado entre ellos. En el extremo final del cuello se encuentra el cátodo, un electrodo que se calienta hasta una temperatura cercana al rojo caliente mediante una corriente que pasa a través de una bobina de calefacción interna. El cátodo se cubre de una sustancia que da un flujo constante de electrones cuando se calienta, y los otros electrodos orientan este flujo en un estrecho cañón de considerable energía. La mayoría de los electrodos del cañón se mantienen a potenciales (voltajes) constantes, alguno de los cuales son bastante altos (varios miles de voltios). El electrodo de rejilla, puede tener su potencial variable, lo que hace que la intensidad del cañón varíe desde cero hasta su máximo.

El haz de electrones puede focalizarse mediante una lente electrónica, del mismo modo que la luz se puede focalizar mediante una lente de vidrio. En los TRC, la lente electrónica es un campo eléctrico producido por electrodos adicionales, o campos magnéticos producidos por bobinas que rodean el cuello del tubo. El segundo método es el que se utiliza normalmente. El rayo

se focaliza hasta un pequeño punto donde golpea la parte frontal del tubo, (la parte interior de la pantalla).

La cara interior de la pantalla está recubierta de una sustancia fosforescente que brilla cuando el haz de electrones la golpea, y durante algún tiempo después (una fracción de segundo).

Acabamos de ver como el campo magnético se puede usar para focalizar el rayo de electrones; de la misma forma, se puede usar para cambiar la dirección del rayo. Más allá, se montan dos conjuntos de bobinas fuera del tubo, al final del cuello, cerca de su unión con el armazón en forma de embudo. Una corriente en un par de bobinas deflece el rayo horizontalmente; el otro par lo deflece verticalmente. Por esto, usando dos pares de bobinas juntas, puede hacerse que el rayo golpee cualquier punto de la pantalla. Puede recorrerse la pantalla para dibujar cualquier patrón requerido, mediante una estela de brillo sobre el material fosforescente.

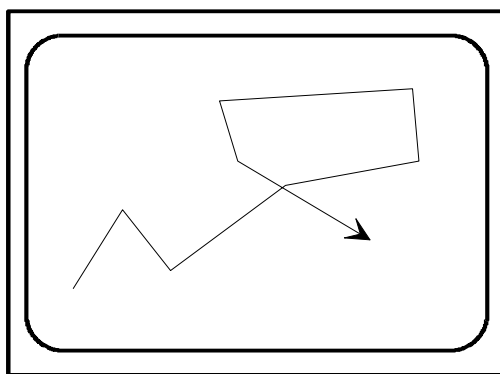


Fig. 5.2 Movimiento típico del haz electrónico sobre un visualizador TRC de barrido aleatorio

En este punto debemos distinguir entre dos métodos para el control del movimiento del rayo (y los puntos de luz resultantes), a través de la pantalla. Estos métodos se conocen como "vector scanning" (barrido aleatorio) que se muestra en la figura (5.2) o "raster scanning" (barrido secuencial) (Fig 5.3). En el barrido aleatorio, la corriente en dos pares de bobinas de deflexión, está bajo el control del computador, y el rayo se puede mover para dibujar cualquier tipo de línea en la pantalla. Este método es el que se usa habitualmente en las pantallas de los osciloscopios analógicos, y necesita tres coordenadas, x , y e intensidad del haz, normalmente denominada z .

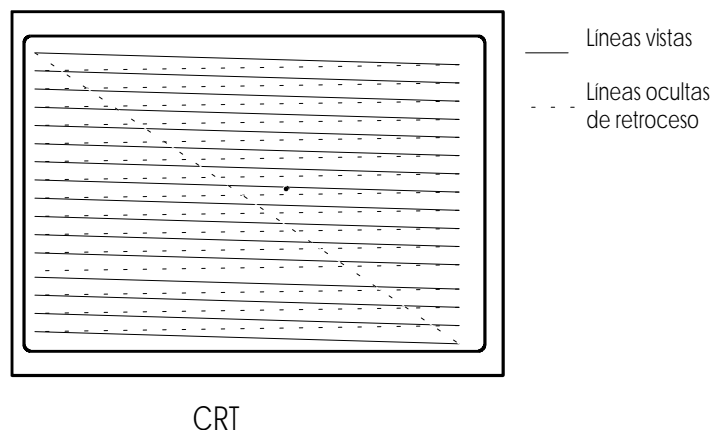


Fig. 5.3 Movimiento del haz en un visualizador TRC de barrido secuencial

El segundo método, de barrido secuencial, se usa por la práctica totalidad de las unidades de vídeo y televisión y va a ser del que nos vamos a ocupar con detalle. Está basado en el dibujo

regular de líneas paralelas igualmente espaciadas en la pantalla. Para hacer esto, la corriente en la bobina de barrido horizontal se va incrementando constantemente. Por lo tanto, el rayo se mueve con una velocidad constante de izquierda a derecha de la pantalla, obteniendo una línea brillante. cuando el rayo encuentra el borde de la pantalla, la corriente se invierte para mover de nuevo el rayo a la parte izquierda de la pantalla; durante el retroceso se usan mayores corrientes para que el rayo se mueva más rápido. Este movimiento inverso es llamado "fly-back" o retrazo horizontal. La red en el cañón de electrones se usa para anular el rayo de electrones durante el retrazo horizontal con objeto de que no se dibuje ninguna línea.

El proceso completo vuelve a repetirse para dibujar otra línea. Sin embargo, la corriente en la bobina de barrido vertical es aumentada también de forma constante, aunque mucho más lentamente, por lo que la segunda línea se dibuja justamente debajo de la primera. El proceso se repite hasta que toda la pantalla se ha llenado con una red de líneas paralelas igualmente espaciadas. La corriente en las bobinas de barrido vertical se invierten entonces para retornar rápidamente al rayo a la parte superior de la pantalla (esto es el retrazo vertical), y de nuevo el rayo se inhibe durante este periodo de retorno (figura 5.3). El proceso completo se repite indefinidamente para dibujar una secuencia idéntica un número elevado de veces por segundo, normalmente entre 60 y 100. La imagen producida en un barrido puede generarse de nuevo, o puede cambiarse en el barrido siguiente; si se genera de nuevo, la rapidez del barrido produce una imagen totalmente estable, y libre de parpadeo ante la visión humana; y si la imagen cambia adecuadamente de un barrido al siguiente, dará la sensación de movimiento continuo.

5.3 ESTUDIO DE UN VISUALIZADOR DE BARRIDO SECUENCIAL

Cualquier visualizador TRC gráfico está constituido fundamentalmente por dos dispositivos: la pantalla del visualizador, y el controlador de pantalla.

5.3.1 La pantalla del visualizador

En los visualizadores TRC de barrido secuencial, la pantalla está constituida por un tubo de rayos catódicos, y por los circuitos que gobiernan el movimiento del haz electrónico, es decir, por un monitor TV.

Un monitor TV responde al diagrama representado en la figura (5.4); en ella, las líneas continuas representan el flujo de la señal que contiene la información de vídeo, las discontinuas el flujo de las señales de sincronismo. El movimiento del haz electrónico sobre la pantalla del TRC, se controla mediante dos tensiones de diente de sierra, generadas en los circuitos osciladores, amplificadas por los circuitos de deflexión, y aplicadas a las bobinas de deflexión horizontal y vertical.

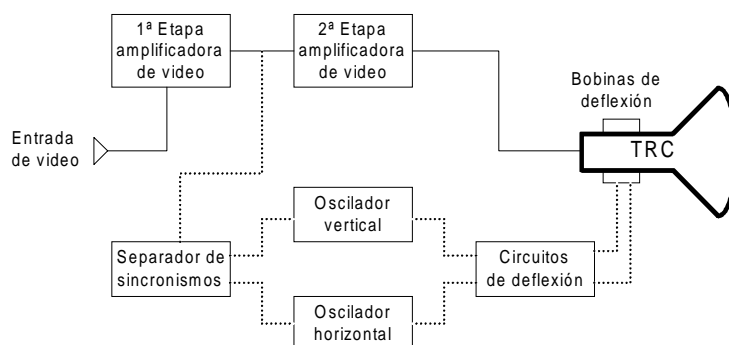


Fig. 5.4 Diagrama de bloques de un monitor de TV

El oscilador horizontal genera una tensión en forma de diente de sierra según la figura (5.5), cuya frecuencia (f_H) depende del tipo de monitor (en los monitores estándares de TV, esta frecuencia es de 15.750 Hz en el sistema americano de TV de 525 líneas y 60 cuadros por segundo, y 15.625 Hz en el sistema europeo de 625 líneas y 50 cuadros por segundo).

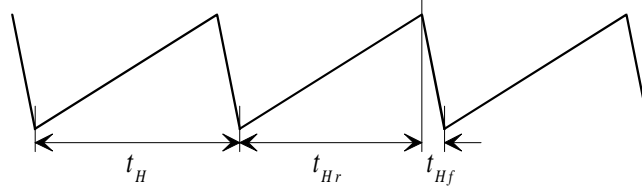


Fig. 5.5 Tensión generada por el oscilador horizontal. Controla el haz electrónico

Esta tensión gobierna el movimiento horizontal o barrido horizontal del haz sobre la pantalla; su frecuencia $f_H = 1/t_H$, se denomina frecuencia de línea. Durante t_{Hr} el haz se mueve desde el borde izquierdo al derecho de la pantalla, y al llegar a este punto, se produce una rápida caída a cero de la tensión ($t_{Hf} \approx 0,20t_H$), que origina el retorno del haz a la posición de partida (Fig. 5.6)

Este movimiento provocaría una única línea horizontal en la pantalla, por lo que para conseguir un barrido de la pantalla completa, es necesario acompañar a este movimiento horizontal de otro vertical. Dicho movimiento se produce por la actuación sobre la bobina de deflexión vertical de una tensión en diente de sierra, generada por el oscilador vertical (Fig. 5.7), la frecuencia de esta tensión (f_V) depende el monitor utilizado, aunque normalmente es cercana a la empleada en la TV comercial, y será de 50 Hz si se utiliza el sistema europeo, o de 60 Hz si el sistema utilizado es el americano. En la actualidad, valores por encima de los 75Hz son muy habituales y algunos monitores profesionales superan los 120Hz. A esta frecuencia vertical se le denomina también frecuencia de refresco ya que indica el número de veces que la pantalla se redibuja en cada segundo. Aunque 50Hz son suficientes para dar un aspecto de imagen estable, se prefieren frecuencias superiores porque producen una menor fatiga visual cuando la vista permanece fija sobre la pantalla durante largos periodos de tiempo y es uno de los factores que más encarecen los monitores.

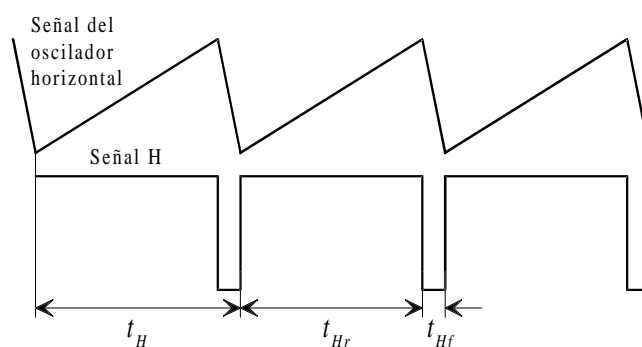


Fig. 5.6 Señal de sincronismo horizontal y su relación con el oscilador horizontal

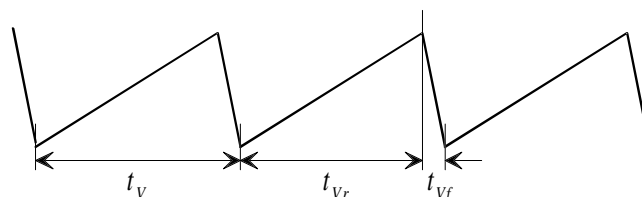


Fig. 5.7 Tensión de control del movimiento vertical del haz electrónico

Durante t_{vf} , el aumento de la tensión aplicada a la bobina de deflexión vertical, produce el movimiento del haz electrónico desde el borde superior al inferior de la pantalla; al llegar al borde inferior, la caída de tensión durante el intervalo t_{vr} ($t_{vf} = 0.08t_v$), provoca el retorno brusco del haz al borde superior, donde comienza de nuevo el proceso.

La combinación de los movimientos horizontal y vertical es el movimiento de barrido del haz sobre la pantalla. El número de líneas horizontales por cada barrido de pantalla viene dado por:

$$n = t_v / t_H$$

De estas n líneas horizontales, no todas son visibles ya que algunas tienen lugar durante el retorno vertical. El número de líneas visibles es:

$$n_r = t_{vr} / t_H$$

y el número de líneas invisibles:

$$n_f = t_{vf} / t_H$$

Si $t_{vf} = 0.08t_v$, obtenemos que: $n_f = 0.08n$.

Además de las dos señales descritas, que son generadas por el propio monitor; el funcionamiento del monitor necesita una sincronización entre los osciladores y el dispositivo que suministra la información a visualizar. La intensidad del haz electrónico se modula de acuerdo con la información a visualizar. Son necesarias por lo tanto, tres señales externas de control, que describiremos a continuación (Fig. 5.8). Recuérdese que la señal H también es un diente de sierra pero como su frecuencia es mucho mayor que la de la señal V se ve como una secuencia de pulsos estrechos.

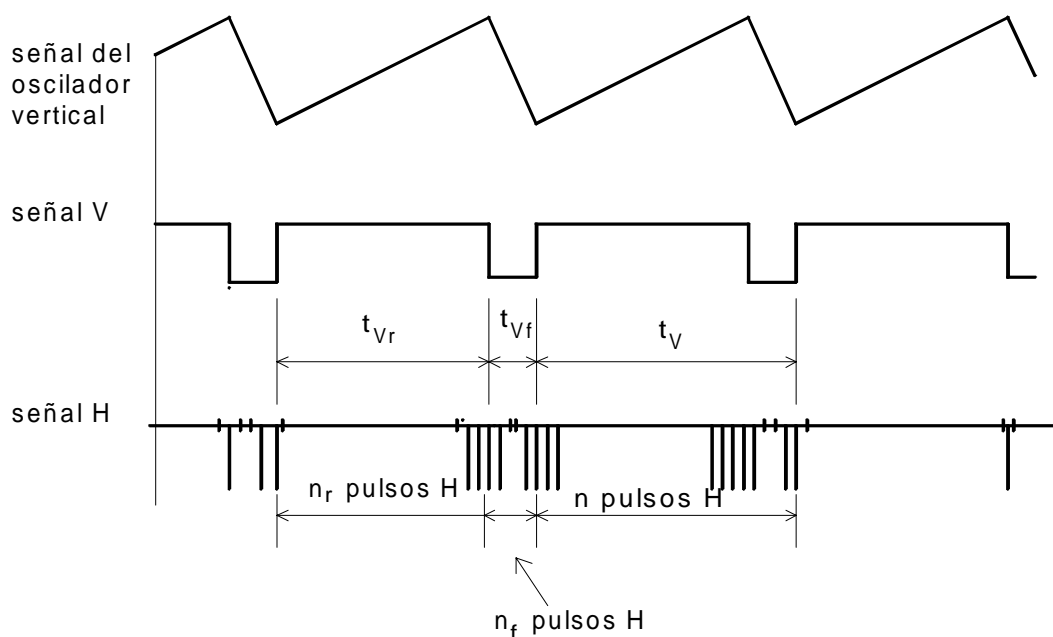


Fig. 5.8 Relación entre los sincronismos horizontal, vertical y oscilador vertical

5.3.2 Sincronismo horizontal o señal H

Es una señal de impulsos, de frecuencia f_H , cuyo fin es sincronizar el oscilador horizontal con la fuente de información de vídeo. Esta sincronización permite a dicha fuente conocer en cada momento la línea horizontal que barre el haz, y así suministrar la información que corresponde a esa línea.

El pulso de sincronismo horizontal determina el inicio y la duración de la caída de tensión generada por el oscilador horizontal.

5.3.3 Sincronismo vertical o señal V

Es también una señal de impulsos, de frecuencia f_V , y permite la sincronización entre la fuente de información (la tarjeta controladora del ordenador) y el oscilador vertical que se regenera en el monitor.

5.3.4 Señal de modulación de la intensidad del haz o señal Z

Simultáneamente con el barrido de la pantalla por el haz electrónico, es necesario un control sobre la intensidad de éste para producir la combinación de tonalidades que constituye la imagen de vídeo. Este control lo realiza la señal Z, que el dispositivo controlador suministra al monitor, y que éste amplifica por medio de etapas amplificadoras de vídeo para aplicarlas al CRT (Fig. 5.9).

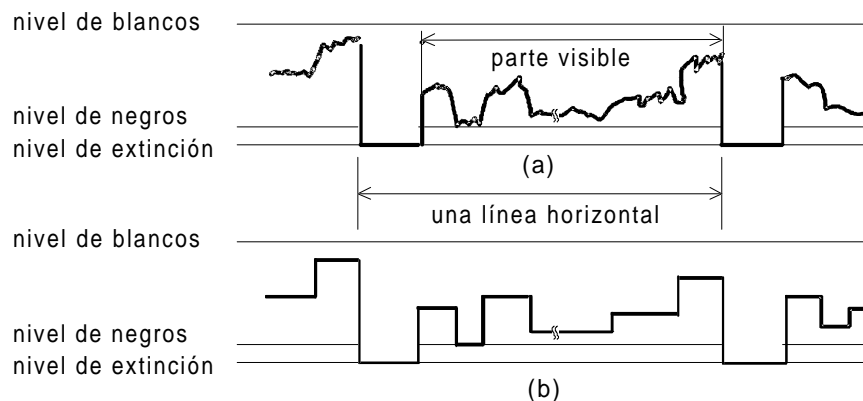


Fig. 5.9 Variación de Z en un barrido horizontal a) TV, b) métodos de procesamiento digital

La tonalidad que el haz electrónico produce al incidir sobre la pantalla, viene determinada por la tensión presente en ese momento en la señal Z. Esta tensión varía entre dos niveles fundamentales, que se denominan 'nivel de negros' y 'nivel de blancos'; la variación entre estos dos niveles puede tener lugar de forma continua, como en la TV comercial, o de forma discreta, como en la generación de imágenes mediante procesamiento digital.

Además de estos dos niveles de referencia, existe un tercer nivel denominado de extinción, durante el cual no se produce traza del haz sobre la pantalla. Este nivel se genera para extinguir la traza de retorno, tanto horizontal como vertical.

En la práctica, estas tres señales descritas no se aplican al monitor separadamente. Mediante una adecuada combinación de ellas se forma una señal, denominada SEÑAL COMPUESTA DE VÍDEO (Fig. 5.10) que gobierna el monitor. En los monitores de color se necesitan además de la señal compuesta de vídeo, las tres señales correspondientes a los tres colores básicos. Es decir, la señal Z deberá contener, codificados de alguna manera las señales correspondientes a los tres colores básicos R (rojo), G (verde) y B (azul). En la televisión, todas estas señales R, G, B y sincronismos se codifican en una sola señal denominada señal de vídeo compuesto. Hay distintas formas de codificar estas señales en una sola dando lugar a los distintos formatos de televisión existentes: PAL, SECAM ó NTSC.

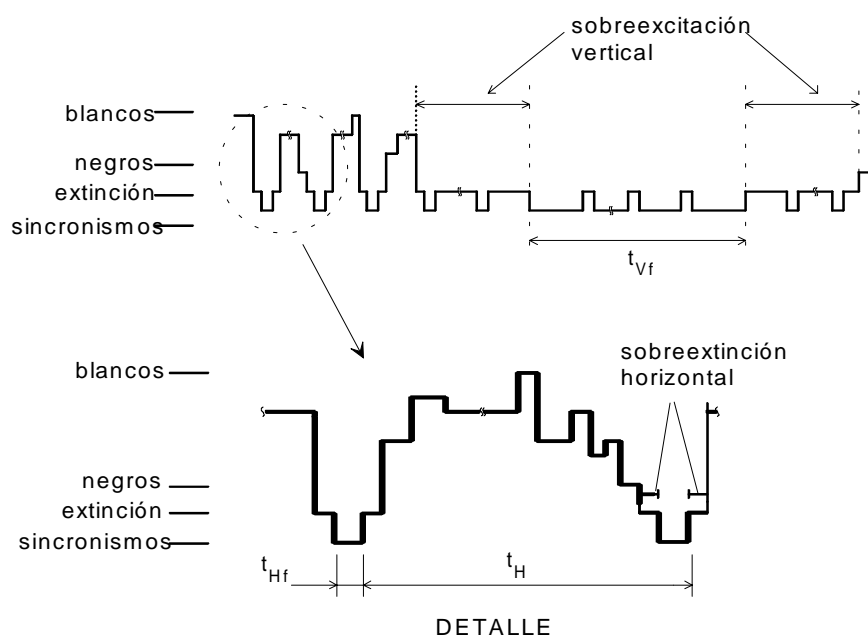


Fig. 5.10 Señal compuesta de vídeo, contiene información de vídeo y los sincronismos

Por el contrario en sistemas informáticos, y debido a que la distancia que separa la fuente de señal del monitor es relativamente pequeña, todas estas señales se envían por separado. Las dos señales de sincronismo se envían en modo simple y la información de vídeo en modo diferencial, por lo que, para un monitor de color, se necesitan un mínimo de 9 hilos (2 por cada uno de los tres colores básicos, 2 de sincronismo y una tierra común para los sincronismos).

Para aislar los sincronismos de la información de vídeo se utiliza el circuito separador de sincronismos. En la figura se observa una sobreextinción en los retornos horizontal y vertical. Esta sobreextinción consta de algunas líneas horizontales (el número depende del dispositivo controlador utilizado), antes y después del retorno vertical; y de una pequeña fracción del barrido horizontal; y su objeto es evitar pérdidas de información en los bordes de la pantalla.

Si el monitor es monocromo, en cada celdilla que corresponde a un punto en la pantalla, la señal de vídeo puede tener dos valores, uno que corresponde al nivel de blanco o punto iluminado y otro se corresponde con el nivel de negro o punto apagado.

En la mayoría de los monitores, existen dos mandos de ajuste desde el exterior, que son el brillo y el contraste. El mando de brillo actúa sobre el amplificador de vídeo y aumenta o disminuye la señal entera. El mando de contraste actúa sólo sobre la parte de la señal de vídeo que está por encima del nivel de negro; es decir, aumenta o disminuye la diferencia entre los niveles de blanco y negro. Los monitores más modernos añaden numerosos controles que tienen como objetivo actuar sobre las señales de sincronismo de forma que la imagen se puede desplazar ligeramente sobre la pantalla, cambiar su tamaño, inclinación, etc.

Veamos un ejemplo de cómo se forma un carácter en la pantalla. Supongamos que queremos representar una 'B' al comienzo de una línea. Para comprender mejor el proceso, puede suponerse que la pantalla está idealmente dividida en celdillas de puntos. Estas celdillas vienen determinadas por la resolución, así una señal de vídeo de 800x600 puntos de resolución, nos da una serie de 800 celdillas por cada una de las 600 líneas horizontales.

Supongamos también que el carácter se forma a partir de una matriz de 7x9 puntos, dejando el espacio de un punto entre dos caracteres consecutivos horizontalmente. Como hemos supuesto que el carácter a representar está en la primera posición de una línea, la señal de vídeo que forma

el carácter está precedida de un sincronismo horizontal. Tal señal está representada en la figura (5.11)

5.3.5 Magnitudes significativas

La figura (5.12) está descompuesta en dos partes. En la primera se observa con detalle que sucede en cada línea, es decir, entre dos sincronismos horizontales; y en la segunda se muestra una visión global de lo que sucede en la pantalla completa. En la tabla (5.1) se puede observar un ejemplo con los valores de las magnitudes que aparecen en la figura.

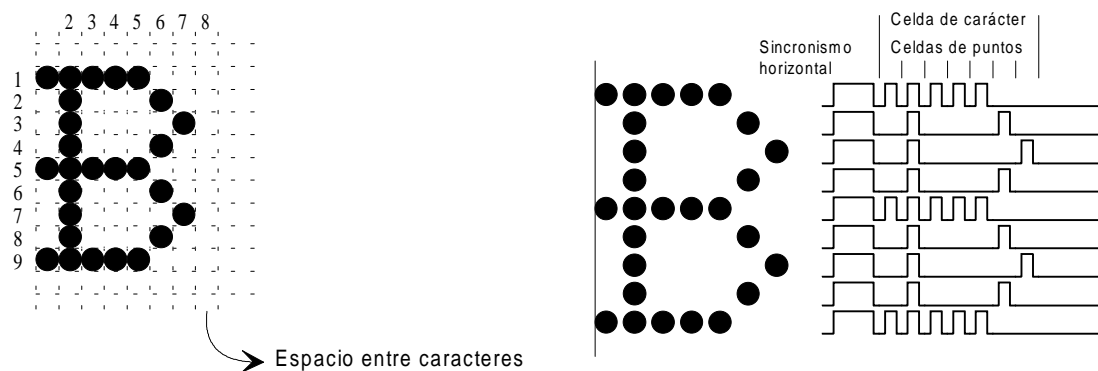


Fig. 5.11 Formación del carácter 'B' en la pantalla

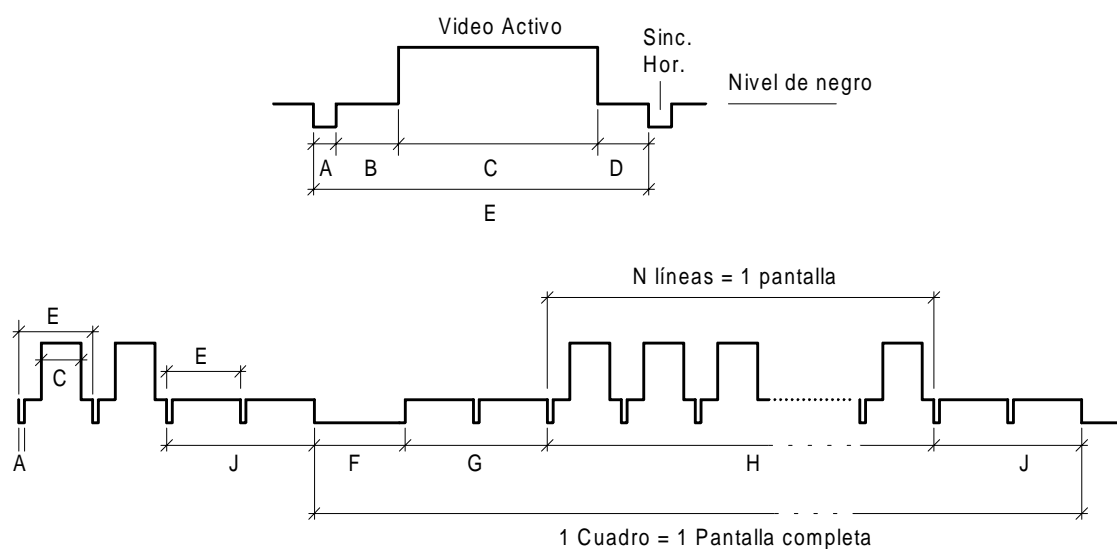


Fig. 5.12 Señal compuesta de vídeo. a) Detalle de una línea. b) Pantalla completa.

Magnitud	Valor típico	Significado
A	5.4 μs	Anchura de pulso de sincronismo horizontal. Tiempo que tarda el haz en retornar
B	8 μs	Intervalo horizontal inicial. Tiempo que va desde el retorno del haz hasta vídeo activo
C	44.7 μs	Tiempo de vídeo activo, en el cual se ve la imagen en pantalla
D	4.4 μs	Intervalo horizontal final. Tiempo que va desde que la imagen desaparece por la derecha hasta que el haz empieza a retornar a la izquierda.
E	62.6 μs	Tiempo total de la línea horizontal. Es la inversa del ancho de banda
F	0.19 ms	Anchura del pulso de sincronismo vertical. Tiempo efectivo de retorno vertical
G	1.82 ms	Intervalo vertical inicial. Tiempo que va desde que el haz comienza su recorrido descendente hasta la primera línea de vídeo visible
H	13.4 ms	Tiempo de vídeo visible por pantalla
J	0.87 ms	Intervalo vertical final. Tiempo que va desde que el haz desaparece por debajo hasta que empieza a subir
K	16.3 ms	Tiempo total efectivo de una pantalla completa

Tabla 5.1 Magnitudes características de la señal de vídeo compuesto

5.4 TIPOS DE MONITORES

Existen varias clasificaciones que se pueden hacer en los monitores, atendiendo a su ancho de banda, al tipo de señal de vídeo que soporten, etc. Comentaremos algunas de estas, que por cierto, no son mutuamente excluyentes.

5.4.1 Monitores mono y multi-frecuencia

Según la frecuencia de barrido horizontal se pueden clasificar los monitores en los que son monofrecuencia y multifrecuencia. Los monofrecuencia tienen un barrido horizontal en el que la distancia entre sincronismos horizontales es fija. Esta distancia dará una idea de la resolución máxima horizontal que soporta. Si la controladora de vídeo le envía una frecuencia distinta la imagen en pantalla no se logrará estabilizar y no se verá nada.

Los multifrecuencia tienen un rango de valores entre los que puede variar el sincronismo horizontal. En estos monitores, el sincronismo horizontal interno que gobierna el haz se consigue, al igual que en las comunicaciones serie o en los dispositivos de almacenamiento, por medio de un PLL ('Phase Latch Loop' o lazo de enganche de fase) interno alimentado por el sincronismo horizontal externo. El PLL tiene un margen de captura muy amplio, lo que hace que la frecuencia de sincronismo horizontal externa pueda variar bastante y aún así el PLL interno sincronizará con ella.

La ventaja de los monitores multifrecuencia radica en que pueden conectarse a una amplia gama de controladores, todos los que caigan dentro del margen de captura. Por ejemplo, un monitor que tenga un rango de captura horizontal desde 15.5 kHz hasta 36 kHz podrá soportar una resolución desde 640x200 hasta 720x480 sin necesidad de ningún ajuste externo.

Algunos monitores pueden sincronizar a distintas frecuencias, dentro de un conjunto de frecuencias predefinido, pero sólo se aplica el calificativo de multifrecuencia a aquellos que pueden sincronizar cualquier frecuencia dentro de un margen dado.

5.4.2 Monitores analógicos y digitales

Hasta ahora hemos supuesto señales de vídeo son generalmente señales digitales. En algunos casos, si el vídeo tiene carácter analógico todas las discusiones anteriores siguen siendo válidas, cambiando los colores por niveles de grises. En los monitores analógicos, la señal de vídeo puede tomar cualquier valor entre el nivel de negro y el de blanco. Esto adquiere especial significado en los monitores de color como veremos en el punto 5.4.4.

5.4.3 Entrelazado

Algunos monitores de gama baja, debido a que tienen un ancho de banda reducido, sólo presentan la mitad de las líneas en cada cuadro, dejando para el siguiente la otra mitad. Esto les permite llegar a mayores resoluciones con el mismo ancho de banda, a costa de reducir a la mitad la frecuencia real de refresco. A este tipo de monitores se les conoce como entrelazados. Hay que tener en cuenta que para que funcionen correctamente, la fuente de vídeo debe proporcionar la señal entrelazada.

5.4.4 Monitores de color

Los monitores en color sólo tienen una diferencia básica con los monocromo. En vez de tener un cañón de electrones tienen tres: uno para el color rojo, uno para el verde y otro para el azul; cada uno de ellos lanza un haz de electrones que chocan en la pantalla con puntos de fósforo de distintos tipos. Cada tipo de fósforo emite un color de la luz al ser excitado por los electrones. El material que recubre la pantalla es sensible a estos cañones de forma que producen el color básico si es bombardeado por uno de ellos o produce la composición si es bombardeado por varios a la vez.

En cuanto a los sincronismos, las discusiones anteriores siguen siendo válidas. La peculiaridad principal es que ahora la señal de vídeo está descompuesta en tres: la señal de rojo (R), la señal de verde (G) y la señal de azul (B).

Existen básicamente dos formas de generar la gama de colores en un monitor. Podemos distinguirlas como formas de gama fija y formas de gama variable.

La forma de gama fija consiste en que las señales básicas de generación de color son digitales, es decir, toman valores de 0 ó 1 en cada punto de la pantalla. Con este esquema, si disponemos sólo de tres señales básicas RGB podremos formar en cada punto de la pantalla una gama de 8 colores distintos, que corresponden a la combinación de los tres cañones de electrones.

La tabla (5.2) muestran los colores obtenidos. En ella un 0 indica que el cañón correspondiente está inactivo y un 1 significa que está activo.

R	G	B	Color
0	0	0	Negro
0	0	1	Azul
0	1	0	Verde
0	1	1	Cíen (Azul cielo)
1	0	0	Rojo
1	0	1	Magenta
1	1	0	Amarillo
1	1	1	Blanco

Tabla 5.2 Formación de colores.

Una posibilidad de aumentar esta gama de colores consiste en añadir una señal extra de intensidad. Así tenemos la posibilidad de la codificación de 4 bits con este esquema y obtenemos una gama de 16 colores.

Una segunda posibilidad para generar una gama fija de colores pero más amplia es descomponer las tres señales de color en seis, dos de cada color. Así la interface dispondría de los sincronismos y de las señales R, G, B, R', G', B'. Con esta nueva estructura se obtiene una gama de 64 colores 2^6 , puesto que tenemos 6 señales digitales para codificar.

Finalmente, la solución más idónea es hacer que el monitor sea analógico. De esta forma la circuitería interna del monitor es sensible a cualquier nivel de tensión que le venga por las tres señales básicas. Este nivel será transmitido proporcionalmente a los cañones para que generen la composición en la pantalla. La composición así obtenida genera en teoría infinitos colores.

En la práctica, lo que se hace es dividir el nivel de continua analógico de las señales básicas en niveles cuantizados que nos dará una gama de colores limitada pero muy elevada. Debe tenerse en cuenta, que el ojo humano tiene una capacidad limitada para distinguir matices de colores, y por tanto, sería inútil disponer de una gama tan amplia en la que no se pudiesen distinguir los colores.

Si por ejemplo, cuantizamos el nivel de las señales RGB a 16 valores posibles dentro de los niveles TTL, obtenemos una gama de colores de:

$$N^{\circ} \text{ de colores} = 2^4 \cdot 2^4 \cdot 2^4 = 2^{12} = 4096 \text{ colores distintos}$$

Esto tiene la ventaja de que un mismo monitor, al admitir cualquier nivel de señal, puede ser excitado por cualquier fuente de vídeo (en cuanto a colores se refiere) con la única limitación de los márgenes de tensión a manejar. Esta es la razón por la que actualmente todos los monitores son analógicos. Actualmente suelen emplearse 8 bits de resolución para cada color, lo que proporciona 256 niveles para cada uno y un total de $256 \times 256 \times 256 = 16.777.216$ colores distintos. Si tenemos en cuenta que un ojo humano entrenado sólo puede distinguir unos 4 millones de colores parece más que suficiente.

5.5 CONTROLADOR DE PANTALLA

En un visualizador gráfico, el controlador de pantalla (Fig. 5.13) es el dispositivo que tiene como misión transformar la información digital resultante del procesado, en las señales que gobiernan la pantalla.

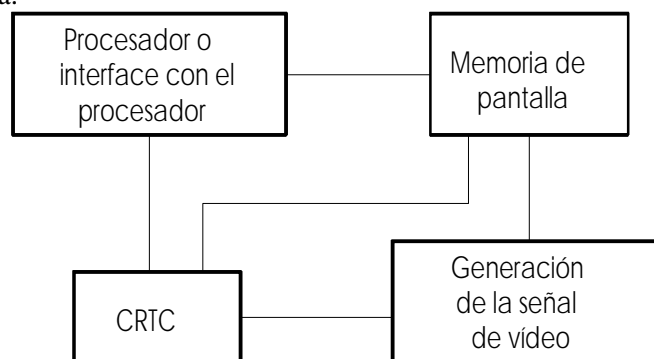


Fig. 5.13 Diagrama de bloques de un controlador de pantalla

Debido a que los monitores no presentan capacidad inherente de memoria, la imagen debe ser generada para cada barrido, es decir, se necesita una regeneración de la imagen a la frecuencia vertical. Esta continua regeneración de la imagen, que se denomina 'refresco de vídeo', requiere un continuo flujo de datos entre el procesador digital y el controlador de monitor. Esto supone una dedicación plena del procesador a la gestión del monitor.

Para evitar esta continua dedicación, el controlador del monitor está formado por dos dispositivos: una memoria digital, 'memoria de refresco' o 'memoria de pantalla' donde el

procesador digital almacena en forma de codificación binaria la imagen que se presentará en la pantalla del monitor, y un procesador gráfico, que leerá el contenido de la memoria y lo transformará en señal de vídeo (señal compuesta de vídeo).

Por lo general, las tarjetas de vídeo disponen de un circuito VLSI ('Very Long Scale Integration') sobre el que recaen las principales funciones de control y que será el que haga las funciones de procesado. A este circuito lo llamaremos CRTC.

5.5.1 Memoria de pantalla

La memoria de pantalla tiene la función de almacenar la codificación binaria de la información correspondiente a uno o varios barridos de la pantalla. El procesador digital grabará la memoria con la codificación correspondiente a un barrido, y el procesador gráfico la leerá secuencialmente a medida que la necesite para generar la información serie de la señal compuesta de vídeo.

Existen tres aspectos fundamentales de la imagen que determinan la capacidad de la memoria de refresco. Estas son:

- Resolución
- Niveles de gris
- Formas de representación

Resolución

Como ya hemos dicho, las imágenes obtenidas por métodos de procesado digital no son continuos, es decir, se obtienen como una adecuada combinación de un número finito de puntos o elementos de la imagen.

Se denomina resolución al número total de elementos que forman la imagen. Cada uno de estos elementos se denomina 'pixel'. La resolución se expresa como NxM pixels, lo que significa que la imagen se ha obtenido utilizando una matriz de N elementos de ancho y M de alto.

El aumento de la resolución aumenta la calidad de imagen, pero también aumenta notablemente la capacidad de memoria de pantalla necesaria.

Niveles de gris

Cada pixel de que consta la imagen puede adoptar distintas tonalidades. El número de tonalidades posibles por pixel se denomina 'niveles de gris' (G), normalmente se dan como potencias de 2: $G = 2^n$. Al igual que ocurre con la resolución, el aumento del número de niveles de gris mejora la calidad de la imagen, pero también aumenta la capacidad de la memoria de pantalla.

Formas de representación

Existen varios métodos de transformar los datos almacenados en la memoria de pantalla de una imagen efectiva de vídeo.

El método más directo consiste en transmitir al monitor, bit a bit, la información almacenada en la memoria de pantalla. Cada '1' lógico se traducirá en un punto luminoso en la pantalla, y cada '0' lógico en un punto oscuro. Así por cada bit de la memoria de refresco existirá un punto en la pantalla del monitor. Un byte de dicha memoria se transforma, mediante una conversión paralelo-serie, en ocho puntos consecutivos en una línea de barrido horizontal. Esta forma de presentación corresponde a una representación puramente gráfica, y la denominaremos 'procesado a bit'. Esta

forma de representación no permite la representación de imágenes con más de dos tonalidades. Para aumentar el número de tonalidades posibles, se asocian varios bits de la memoria de pantalla a un solo punto de la pantalla. Para alcanzar G niveles de gris se necesitan n bits de la memoria por cada punto de la pantalla, donde n viene dado por la ecuación anterior ($G = 2^n$). Esta forma de representación la denominaremos 'mapeado a n bits'

Los métodos de representación mediante mapeado tiene una gran versatilidad, pero también algunos inconvenientes:

- ♦ La gran capacidad de memoria de pantalla que se necesita para almacenar la información correspondiente a un barrido vertical de la pantalla. Por ejemplo, para la representación de una imagen con una resolución de 1024x768 puntos y con 16 niveles de gris (mapeado a 8 bits), se necesitan 6.291.496 bits (768 kB) de memoria de pantalla.
- ♦ Consecuencia de la elevada capacidad de la memoria de pantalla es el tiempo necesario para la representación.

La presentación en pantalla de letras y números únicamente, requiere una capacidad de memoria mucho menor. Una letra de un solo color sobre un fondo uniforme oscuro que ocupe ocho filas de ocho puntos cada una (matriz 8x8) necesita ocho bytes (64 bits) de memoria si la representamos mediante mapeado de bit, pero podemos codificar dicha letra en ASCII y solo necesitará un byte de memoria. Esto significa que en lugar de transmitir cada dato al monitor, es necesario decodificar cada carácter almacenado en la memoria de refresco y generar adecuadamente la información de vídeo. Este proceso lo realiza un dispositivo denominado 'generador de caracteres'.

La utilización de un byte por cada carácter hace que puedan generarse 256 caracteres distintos: un juego completo de mayúsculas, minúsculas, números, signos de puntuación y caracteres especiales. Esta forma de representación se denomina alfanumérica. No obstante, salvo en sistemas muy simples, ha dejado de emplearse debido fundamentalmente a dos razones:

Por una parte, la tecnología ha permitido abaratar las memorias y acelerar su acceso, atenuando los problemas originales, y por otra parte, los nuevos entornos de trabajo de naturaleza puramente gráfica resulta incompatible con los modos alfanuméricos.

5.5.2 El procesador gráfico

Este dispositivo, al igual que la memoria de pantalla, forma parte del controlador de pantalla, y su misión es la lectura de la información binaria almacenada en la memoria para transformarla en la señal de vídeo. El texto (lo llamamos de esta forma aunque la información sea gráfica) debe ser generado carácter a carácter y línea a línea. Una vez terminada su reproducción en la pantalla, debe comenzarse nuevamente para obtener una imagen perfectamente estable.

Los caracteres a reproducir se hallan almacenados en la memoria de pantalla utilizando el código ASCII, o cualquier otro código: el código correspondiente a cada carácter debe ser leído por el procesador gráfico y convertido a señal de información de vídeo, para luego superponer a esta señal los impulsos de sincronismo y obtener la señal compuesta de vídeo.

Como consecuencia de lo expuesto, se deduce que un procesador gráfico estará compuesto por dos partes fundamentales:

- El generador de la señal de información de vídeo.
- El generador de las señales de control y temporización.

5.6 GENERACIÓN DE LA SEÑAL DE VIDEO

5.6.1 Generador de caracteres

Para representar los caracteres en la pantalla del monitor se utiliza un formato basado en una matriz de puntos (en lo sucesivo nos referiremos al formato 5x7). La información correspondiente a las matrices de puntos de los distintos caracteres, que admitan representación en el procesador gráfico utilizado, está almacenada en una memoria ROM que se denomina 'generador de caracteres' (Fig. 5.14). A la entrada del generador de caracteres, como direccionamiento de la memoria, se presenta el código del carácter y la fila de la matriz que corresponde a la línea actual del barrido horizontal; a la salida, como dato de la memoria, obtenemos en paralelo la información correspondiente a dicha fila.

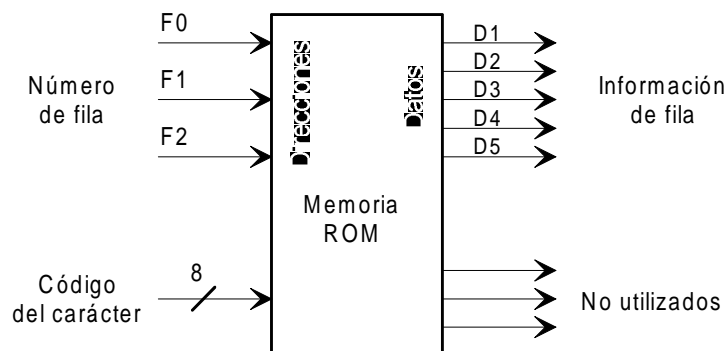


Fig. 5.14 Esquema de un generador de caracteres

Veamos a continuación como se puede representar una línea de texto formada por q caracteres. Si utilizamos una matriz de puntos de formato 5x7, necesitamos como mínimo 8 barridos horizontales de la pantalla, 1 barrido por cada fila de la matriz y otro, como mínimo, para conseguir la separación entre dos líneas de caracteres.

En el primer barrido se representan los q grupos de 5 puntos de los q caracteres de la línea que corresponden a la primera fila de la matriz de puntos; en el segundo barrido se hace lo mismo para la segunda fila y así sucesivamente hasta conseguir la representación de las 7 filas de la matriz de puntos.

Para representar una línea completa de caracteres se van leyendo de la memoria los códigos de éstos en forma secuencial, tantas veces como filas tiene la matriz de puntos; es decir, deben leerse los códigos de todos los caracteres en forma secuencial en cada barrido horizontal. Es necesario, por tanto, un contador que nos indique cual es el número de carácter de la línea que se está representando en aquel momento, y otro contador que nos indique cual es el número de la fila de la matriz de puntos que debemos visualizar.

Como se dispone también de un contador de líneas de texto, en todo momento está determinada la dirección absoluta de la posición de memoria que se debe leer en cada momento para obtener el código del carácter a visualizar. Este código se introduce en el generador de caracteres y, mediante tres líneas de control, se selecciona la fila de la matriz de puntos que debe visualizarse. A la salida tenemos en paralelo la información correspondiente a la fila de la matriz de puntos direccionada, información que se carga en paralelo en un registro de desplazamiento gobernado mediante una señal de reloj, cuya frecuencia coincide con la frecuencia de visualización de los puntos en la pantalla (Fig. 5.15). Este registro de desplazamiento es de 8 bits, tres más que salidas tiene el generador, y que corresponden a los espacios de separación entre dos caracteres consecutivos, por lo que se encuentran fijados a nivel lógico bajo.

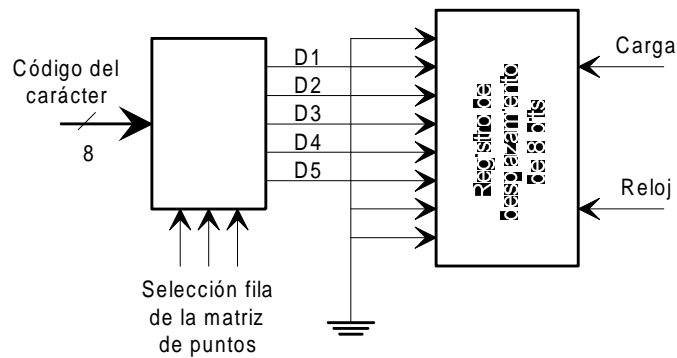


Fig. 5.15 Generador de caracteres con registro de desplazamiento

En la generación de una línea de texto deberá tenerse en cuenta el tiempo de extinción horizontal, durante el cual debe suspenderse la presentación, para reiniciarla en una nueva línea poniendo a cero el contador del número de carácter de línea y repitiendo nuevamente el proceso.

La generación de una página de texto que contenga 8 líneas de carácter se realiza repitiendo 8 veces el proceso descrito. Al igual que durante la extinción horizontal, durante la extinción vertical se actuará suspendiendo la presentación (y por tanto la lectura de memoria), y poniendo a cero el contador de líneas para comenzar un nuevo barrido vertical.

El generador de señales de control y temporización

Este generador, normalmente denominado 'generador de sincronismos', produce señales H y V de control del monitor junto con señales adicionales que controlan de forma temporal los contadores y registros asociados al procesador gráfico.

El generador de sincronismos está basado en un oscilador cuya frecuencia debe ser múltiplo simultáneamente de la frecuencia de línea (f_H), de la frecuencia de cuadro (f_V) y de la frecuencia de visualización de puntos, ya que están relacionadas a través de números enteros, es decir, cada línea tiene un número entero de puntos, y cada pantalla tiene un número entero de líneas.

Las frecuencias de línea y de cuadro dependen del tipo de monitor utilizado y están relacionadas por:

$$f_V = f_H / M'$$

donde M' representa la máxima resolución vertical, que coincide con el número de líneas horizontales, n_r .

Calculemos a continuación la frecuencia fundamental del generador de sincronismos. El período entre dos puntos será:

$$T = \frac{t_{Hr} \mu\text{seg} / \text{linea}}{N' \text{ puntos} / \text{linea}}$$

donde:

t_{Hr} = tiempo efectivo de barrido horizontal (excluido el retorno)

N' = N° total de puntos en una línea

Si consideramos por ejemplo que $t_{Hf} = 0.2t_H \Rightarrow t_{Hr} = t_H - t_{Hf} = 0.8t_H$

Así tenemos que: $T = \frac{0.8t_H}{N'} \mu seg$

Con lo que la frecuencia del oscilador será:

$$f_{osc} = \frac{1}{T} = \frac{N'}{0.8t_H} = \frac{N' f_H}{0.8}$$

y si la expresamos en función de f_V :

$$f_{osc} = \frac{N' M'}{0.8} f_V MHz$$

A partir de la frecuencia base se obtienen las frecuencias necesarias:

$$f_{puntos} = f_{osc}$$

$$f_{carácter} = f_{osc} / p$$

$$f_{lin} = f_H = M' f_V$$

$$f_{cuadro} = f_V$$

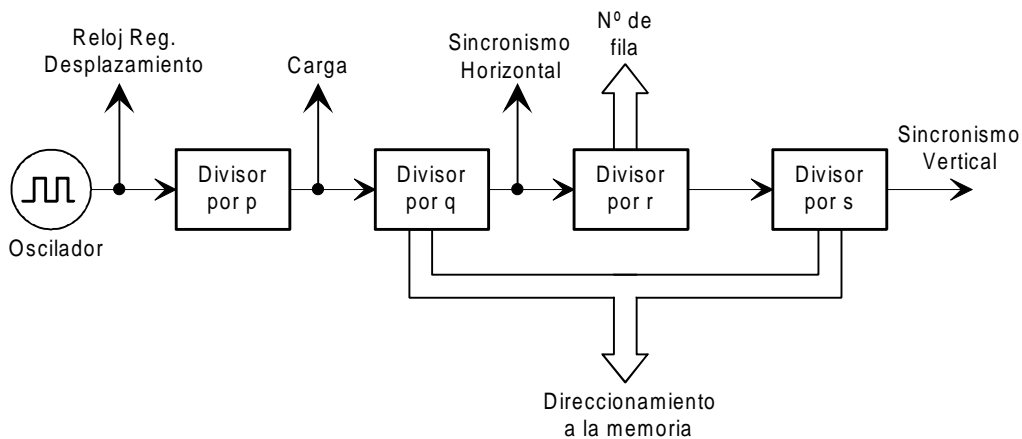


Fig. 5.16 Composición del circuito generador de sincronismos en un procesador gráfico

La forma de construir el generador de sincronismo consiste en ir añadiendo en cascada, a partir del oscilador principal diversos divisores que nos vayan dando las frecuencias necesarias (Fig. 5.16). Los contadores y duración de los impulsos de sincronismo pueden realizarse a partir de las salidas de los divisores y de una pequeña circuitería lógica adicional.

La conexión del circuito generador de sincronismo con el generador de información de vídeo y con la memoria de refresco se muestra en la figura 5.17.

p = nº de puntos por carácter

q = nº de caracteres por línea

r = nº de barridos por línea de caracteres

s = nº de líneas de caracteres por cada pantalla

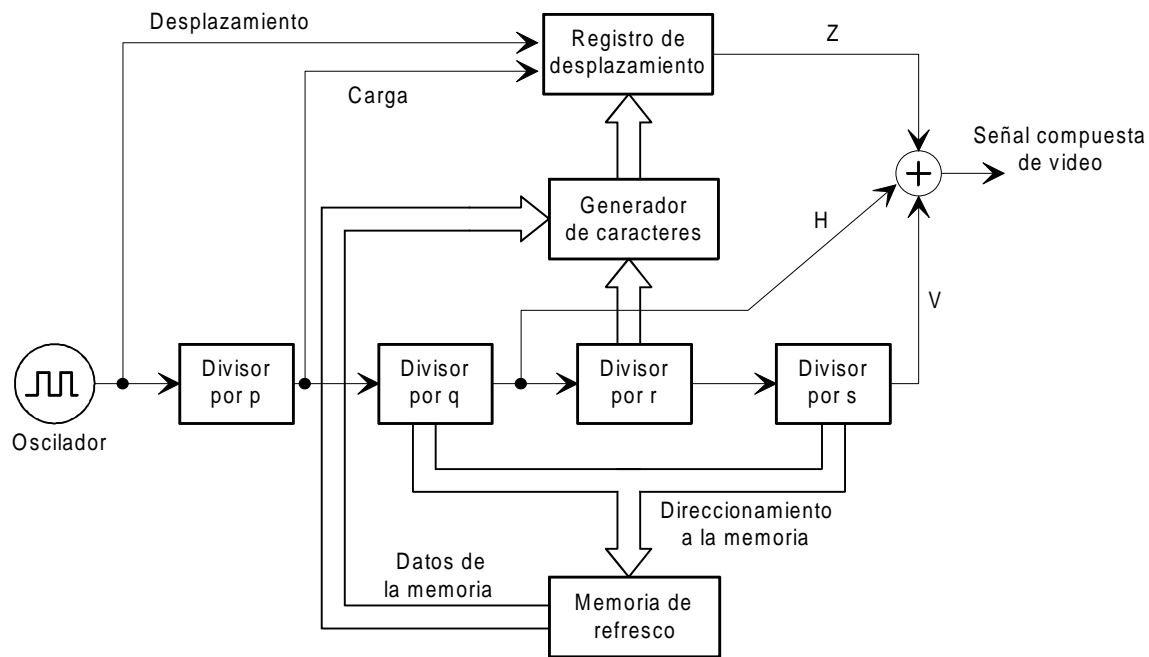


Fig. 5.17 Diagrama de bloques de un circuito controlador de pantalla alfanumérico

Los contadores y registros asociados a la cadena de divisores proporcionan, en forma implícita, las coordenadas del haz sobre la pantalla en cualquier momento, y la dirección y coordenadas del carácter que se está visualizando.

Como registros básicos podemos mencionar:

Indicadores de las coordenadas del haz electrónico:

- Registro del número de línea de barrido
- Registro de posición del haz dentro de la línea

Indicadores de las coordenadas del carácter visualizado:

- Registro del número de línea de caracteres
- Registro del número de carácter en la línea
- Registro del número de fila de la matriz de puntos
- Registro del número de columna de la matriz

La dirección absoluta en la memoria del carácter a visualizar se obtiene a partir de la dirección del primer carácter del texto a visualizar, del registro de línea de caracteres, y del registro de posición del carácter dentro de la línea.

Todo lo expuesto en este apartado es igualmente válido, aunque con algunas simplificaciones para procesadores gráficos que utilicen la forma de representación del mapeado a bit. Este tipo de procesadores gráficos elimina el generador de caracteres, ya que la correspondencia es directa; la información obtenida de la memoria se presenta directamente del registro de desplazamiento que lo serializa para formar la señal de vídeo, lo que permite la eliminación de algunos registros y contadores.

5.6 EJEMPLOS DE TARJETAS

MDA

La tarjeta MDA fue diseñada para utilizarse con un monitor monocromo de 80 columnas y 25 filas de texto alfanumérico.

- Resolución: 720x350
- Matriz de carácter: 9x14
- Modo gráfico: no tiene

CGA

- Modo texto: 25 filas por 80 columnas
- Matriz de carácter: 8x8
- Modo gráfico: 640x200

HGC, HGC+, EGA, Hercules InColor, MCGA, VGA

Todas las tarjetas tienen partes de su hardware programables, con lo que podríamos controlar la operación de la tarjeta y su presentación en pantalla.

En la figura 5.18 podemos observar los componentes programables del subsistema de vídeo. Estudiaremos cada uno de ellos por separado, comenzando por el buffer de vídeo, que es un bloque RAM donde se almacenan los datos que van a ser presentados por pantalla y que están dentro del mapa de direcciones de la CPU del ordenador.

- Hardware de visualización del color y los caracteres:

Es el hardware adicional el que se encarga de leer y decodificar los datos del buffer de vídeo, como pueden ser el generador de caracteres, el decodificador de atributos, etc.

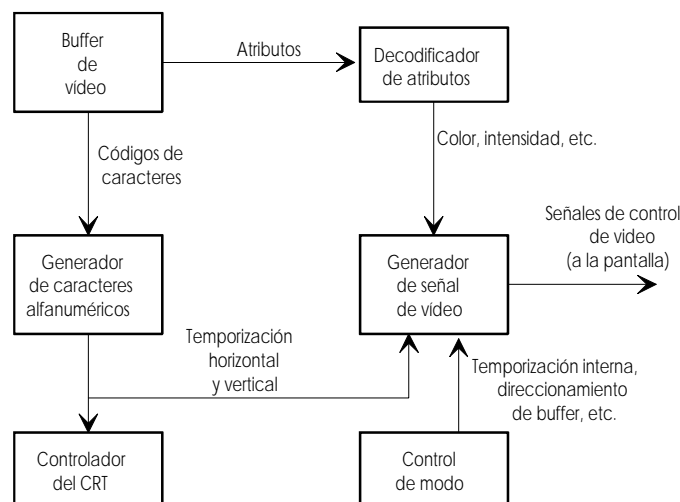


Fig. 5.18 Componentes programables del subsistema de vídeo

- El controlador de CRT (CRTC):

Este dispositivo genera las señales de sincronismo horizontal y vertical. También incrementa un contador de direcciones del buffer de vídeo a una velocidad sincronizada con las señales de barrido. La circuitería de vídeo de la pantalla lee los datos del buffer de vídeo utilizando los valores de las direcciones de CRTC, decodifica los datos y envía las señales resultantes de color y brillo al monitor con las señales de barrido del CRTC. De esta forma el CRTC sincroniza la pantalla de datos desde el buffer de vídeo con las señales de barrido que controlan la pantalla.

El CRTC realiza también otras funciones. Por ejemplo, determinar el tamaño y posición del cursor que se visualiza en la pantalla, seleccionar la parte del buffer de vídeo que se ha de presentar en pantalla, mostrar subrayado, y detectar las señales del lápiz óptico.

En MDA, CGA y tarjetas Hercules, el CRTC es un único chip: el 6845 de Motorola. En EGA el CRTC es un chip LSI diseñado por IBM. En MCGA el CRTC es una parte de su matriz controladora de memoria. Y el CRTC de VGA es un componente del propio chip VGA.

Registro	Nombre	Nombre original	Acceso lectura/escritura
00H	Total horizontal	<i>Horizontal Total</i>	SE
01H	Visualización horizontal	<i>Horizontal Displayed</i>	SE
02H	Posición de sincronismo horizontal	<i>Horizontal Sync Position</i>	SE
03H	Ancho del pulso de sincronismo horizontal	<i>Horizontal Sync Pulse Width</i>	SE
04H	Total vertical	<i>Vertical Total</i>	SE
05H	Ajuste total vertical	<i>Vertical Total Adjust</i>	SE
06H	Visualización vertical	<i>Vertical Displayed</i>	SE
07H	Posición de sincronismo vertical	<i>Vertical Sync Position</i>	SE
08H	Modo entrelazado	<i>Interlace Mode</i>	SE
09H	Línea del barrido máxima	<i>Maximun Scan Line</i>	SE
0AH	Comienzo del cursor	<i>Cursor Start</i>	SE
0BH	Fin del cursor	<i>Cursor End</i>	SE
0CH	Dirección alta de comienzo	<i>Start Address High</i>	SE
0DH	Dirección baja de comienzo	<i>Start Address Low</i>	SE
0EH	Posición alta del cursor	<i>Cursor Location High</i>	L/E
0FH	Posición baja del cursor	<i>Cursor Location Low</i>	L/E
10H	Lápiz óptico alto	<i>Light Pen High</i>	SL
11H	Lápiz óptico bajo	<i>Light Pen Low</i>	SL

Tabla 5.3 Registros de datos del CRTC 6845 de Motorola
(Utilizado en los adaptadores de vídeo: Hércules, MDA y CGA)

5.6.1 PROGRAMACIÓN DEL CONTROLADOR DEL CRT

MDA

El CRTC del adaptador de pantalla monocromo, el 6845 de Motorola, tiene 19 registros internos de datos de 8 bits (tabla 5.3). El contenido de cada registro de control varía con las señales de barrido generadas por el 6845. Uno de estos registros es un registro de direcciones, su contenido indica cual de los otros 18 registros puede ser accedido. La mayoría son sólo de escritura pero en los registros 0EH y 0FH, que controlan la posición del cursor, se puede escribir y leer. En MDA, el registro de direcciones del 6845 está configurado en el puerto de E/S 3B4h y el resto, 18 registros está en el siguiente puerto, de E/S 3B5h.

Para acceder a los registros de datos del 6845 primero se debe escribir el número de registro en el registro de direcciones del 6845 (puerto E/S 3B4h), luego se puede acceder al registro de datos deseado con una lectura o escritura E/S del puerto 3B5h.

CGA

Funciona igual que MDA puesto que también tiene un CRTC 6845 (Tabla 5.3), sin embargo, el registro de direcciones del CRTC está configurado en el puerto E/S 3D4h y se accede a los registros de datos por el puerto 3D5h. Si en un momento dado, queremos saber si nuestro sistema de vídeo es MDA o CGA lo podemos ver en la variable ADDR_6845, que está localizada en la dirección 0040:0063 en el área de datos de vídeo de BIOS.

ADAPTADORES HERCULES

Las tarjetas HGC, HGC+ y la InColor utilizan el 6845 en los puertos E/S en las direcciones 3B4h y 3B5h.

Registro	Nombre	Nombre original	Acceso lectura/escritura
00H	Total horizontal	<i>Horizontal Total</i>	L/E
01H	Visualización horizontal	<i>Horizontal Displayed</i>	L/E
02H	Comienzo del sincronismo horizontal	<i>Start Horizontal Sync</i>	L/E
03H	Ancho del pulso de sincronismo	<i>Sync Pulse Width</i>	L/E
04H	Total vertical	<i>Vertical Total</i>	L/E
05H	Ajuste total vertical	<i>Vertical Total Adjust</i>	L/E
06H	Visualización vertical	<i>Vertical Displayed</i>	L/E
07H	Comienzo del sincronismo vertical	<i>Start Vertical Sync</i>	L/E
08H	(Reservado)		L/E
09H	Líneas de barrido por carácter	<i>Scan Lines Per Character</i>	L/E
0AH	Comienzo del cursor	<i>Cursor Start</i>	L/E
0BH	Fin del cursor	<i>Cursor End</i>	L/E
0CH	Dirección alta de comienzo	<i>Start Address High</i>	L/E
0DH	Dirección baja de comienzo	<i>Start Address Low</i>	L/E
0EH	Posición alta del cursor	<i>Cursor Location High</i>	L/E
0FH	Posición baja del cursor	<i>Cursor Location Low</i>	L/E
10H	Control de modo	<i>Mode Control</i>	L/E
11H	Control de interrupción	<i>Interrupt Control</i>	L/E
12H	Generador de caracteres, polaridad de sincronismo	<i>Character Generator, Sync Polarity</i>	L/E
13H	Puntero al generador de caracteres	<i>Character Generator Pointer</i>	L/E
14H	Contador del generador de caracteres	<i>Character Generator Count</i>	L/E
20-3FH	(Reservado)		L/E

Tabla 5.4 Registros de datos del controlador de memoria MCGA.

Los registros 00H a 0FH son comparables a algunos del controlador de CRT de CGA

EGA

El CRTC de la EGA es un LSI con un conjunto de registros diferentes a los que se utilizan en el 6845 (Tabla 5.4). La interface de programación es similar a la del 6845, pero la asignación de registros y formatos es lo suficientemente diferente como para que los programas escritos para uno y otro tipo de tarjeta no sean compatibles. El CRTC de EGA soporta un conjunto más amplio de funciones de control que el 6845

MCGA

Aquí, las funciones del CRTC están integradas en un componente del circuito denominado Memory Controller Gate Array. Los primeros 16 registros del controlador de memoria son

análogos a los del 6845 (Tabla 5.4). Como en CGA, todos los registros del controlador de memoria MCGA, incluyendo los registros del CRTC, están indexados en un registro de direcciones del puerto E/S ED4h. Puede acceder a los registros de datos a través del puerto 3D5h

Hay algunas características propias del CRTC de MCGA que hacen que se distinga del 6845 de CGA: todos los registros del controlador de memoria son tanto para lectura como para escritura; aún más, los registros 00H hasta 07H, pueden designarse como sólo lectura, de forma que los parámetros de barrido horizontal y vertical no se pueden perder por equivocación. Los registros 00H hasta 07H, se protegen poniendo a 1 el bit del registro de control de modo del controlador de memoria (10H)

Registro	Nombre	Nombre original	Acceso lectura/ escritura
00H	Total horizontal	<i>Horizontal Total</i>	SE
01H	Fin de la activación de pantalla horizontal	<i>Horizontal Display Enable End</i>	SE
02H	Comienzo del blanqueo horizontal	<i>Start Horizontal Blanking</i>	SE
03H	Fin del blanqueo horizontal	<i>End Horizontal Blanking</i>	SE
04H	Comienzo del retrazo horizontal	<i>Start Horizontal Retrace</i>	SE
05H	Fin del retrazo horizontal	<i>End Horizontal Retrace</i>	SE
06H	Total vertical	<i>Vertical Total</i>	SE
07H	Desbordamiento	<i>Overflow</i>	SE
08H	Preselección del barrido de línea	<i>Preset Row Scan</i>	SE
09H	Dirección de la línea de barrido máxima	<i>Maximun Scan Line Address</i>	SE
0AH	Comienzo del cursor	<i>Cursor Start</i>	SE
0BH	Fin del cursor	<i>Cursor End</i>	SE
0CH	Dirección alta de comienzo	<i>Start Address High</i>	L/E
0DH	Dirección baja de comienzo	<i>Start Address Low</i>	L/E
0EH	Posición alta del cursor	<i>Cursor Location High</i>	L/E
0FH	Posición baja del cursor	<i>Cursor Location Low</i>	L/E
10H	Comienzo del retrazo vertical	<i>Vertical Retrace Start</i>	SE
10H	Lápiz óptico alto	<i>Light Pen High</i>	SL
11H	Fin del retrazo vertical	<i>Vertical Retrace End</i>	SE
11H	Lápiz óptico bajo	<i>Light Pen Low</i>	SL
12H	Fin de la activación de pantalla vertical	<i>Vertical Display Enable End</i>	SE
13H	Desplazamiento (ancho línea lógica)	<i>Offset (Logical Line Width)</i>	SE
14H	Posición del subrayado	<i>Underline Location</i>	SE
15H	Comienzo blanqueo vertical	<i>Start Vertical Blanking</i>	SE
16H	Fin blanqueo vertical	<i>End Vertical Blanking</i>	SE
17H	Control de modo	<i>Mode Control</i>	SE
18H	Comparar línea	<i>Line Compare</i>	SE

SE= sólo escritura; SL= sólo lectura; L/E= Lectura/ escritura

Tabla 5.5 Registros de datos del controlador del CRT de EGA y VGA

5.6.1.1 Cálculos elementales del CRTC

Para usar efectivamente el CRTC debe ser capaz de realizar los cálculos elementales necesarios para especificar correctamente las temporizaciones del CRTC; estos cálculos se basan

en tres parámetros: el ancho de banda de la señal de vídeo enviada al monitor y las frecuencias de sincronismo vertical y horizontal del monitor.

Reloj de puntos

Los subsistemas de vídeo del PC visualizan los pixels a una velocidad determinada por el hardware. Esta frecuencia es conocida en el mundo de la electrónica con diferentes nombres: ancho de banda de vídeo, frecuencia de puntos o frecuencia de pixels. El oscilador que genera la frecuencia se denomina reloj de puntos.

La MDA, CGA y Hercules utilizan solo un reloj de puntos. En EGA y VGA, por el contrario, se utilizan varios de estos relojes. Mientras más alta sea la frecuencia del reloj de puntos, mejor será la resolución de la pantalla.

Dado un ancho de banda de vídeo, se puede programar el CRTC de tal forma que las frecuencias de barrido horizontal y vertical enviadas a la pantalla estén limitadas por las frecuencias que puede manejar.

Subsistema IBM	Ancho de banda de vídeo en MHz	Frecuencia de barrido horizontal en KHz	Frecuencia de barrido vertical en Hz
MDA, HGC			
720 X 350 mono	16.257	18.43	50
CGA			
640 X 200 color	14.318	15.75	60
EGA			
640 X 350 color	16.257	21.85	60
640 X 200 color	14.318	15.75	60
720 X 350 mono	16.257	18.43	50
InColor			
720 X 350 color	19.000	21.80	60
MCGA			
640 X 400 mono/ color	25.175	31.50	70
640 X 480 mono/ color	25.175	31.50	60
VGA			
640 X 400 mono/ color	25.175	31.50	70
720 X 400 mono/ color	28.322	31.50	70
640 X 480 mono/ color	25.175	31.50	60
640 X 350 mono/ color	25.175	31.50	70

Tabla 5.6 Barridos básicos para los subsistemas de vídeo de IBM

VGA

Desde el punto de vista funcional, los registros del CRTC de VGA constan de un superconjunto que incluye los registros de CRTC de EGA (Tabla 5.5). El conjunto de registros del CRTC de VGA es direccionable desde los mismos puertos E/S y los de EGA.

El punto más importante es que las especificaciones de los registros del CRTC de EGA han sido llevados al de VGA. Por dicha razón, los programas escritos para los registros de CRTC de EGA pueden funcionar, sin hacer ningún cambio, en el hardware de VGA.

Barrido horizontal

Vamos a ver como se calcularán los valores de los registros típicos de CRTC para una MDA con un monitor monocromo. El ancho de banda de vídeo MDA (velocidad de puntos) es de 16,257 MHz, o lo que es lo mismo: 16.257.000 puntos por segundo. La frecuencia de barrido horizontal de monitor monocromo es de 18,432 (18.432 líneas por segundo). Dividiendo la frecuencia de puntos por la de barrido horizontal se obtienen 882 puntos por línea. Cada carácter visualizado por MDA tiene un ancho de 9 puntos, por tanto, el número total de caracteres en cada línea es 882/9, o lo que es lo mismo, 98.

Este valor se utiliza para programar el registro total horizontal del CRTC. En el caso del CRTC de MDA, 6845 de Motorola, el valor almacenado en este registro debe ser uno menos que el calculado anteriormente, 97 (61h) (Tabla 5.7). Durante este periodo se visualizan realmente 80 caracteres (este es el utilizado por el registro de visualización horizontal, Horizontal Displayed). Las otras 18 frecuencias de caracteres se utilizan en el margen horizontal y en el retraso horizontal.

La duración del intervalo de retraso horizontal es del 10 al 15% del valor del total horizontal. El valor exacto depende del subsistema de vídeo. En MDA, el retraso horizontal se define a 15 caracteres, almacenando dicho valor en el registro de Ancho de sincronismo horizontal del CRTC (Horizontal Sync Width). Esto origina tres caracteres de margen horizontal. Las señales de retraso horizontal se programan para comenzar dos caracteres después de que se visualiza el carácter de más a la derecha al almacenar el valor 82 (52h) en el registro de posición de Sincronismo Horizontal de CRTC; así pues, hay dos caracteres de margen horizontal derecho y uno de margen izquierdo.

Registro	Nombre	Parámetro	Descripción
00H	Total horizontal	97 (61H)	(Total de caracteres por línea barrida)- 1
01H	Visualización horizontal	80 (50H)	Caracteres visualizados en cada línea barrida
02H	Posición de sincronismo horizontal	82 (52H)	Posición donde comienza el retraso horizontal en la línea barrida
03H	Ancho del pulso de sincronismo horizontal	15 (0FH)	Duración del intervalo de retraso horizontal
04H	Total vertical	25 (19H)	Total filas de caracteres en una trama
05H	Ajuste total vertical	2	Restantes filas barridas en una trama
06H	Visualización vertical	25 (19H)	Filas de caracteres visualizadas en cada trama
07H	Posición de sincronismo vertical	25 (19H)	Posición donde comienza el retraso vertical en la línea barrida
08H	Modo entrelazado	2	Siempre a 2
09H	Línea de barrido máxima	13 (0DH)	(Peso de un carácter en líneas de barrido - 1)

Tabla 5.7 Parámetros típicos CRTC para el adaptador de pantalla monocromo (MDA)

Barrido vertical

Al igual que en el caso anterior, se realizan las mismas consideraciones de programación de CRTC para generar el barrido vertical apropiado. La frecuencia de barrido horizontal nominal para

MDA es de 18.432 kHz (18432 líneas por segundo) con una frecuencia vertical de 50 Hz (50 tramos por segundo) de tal forma que el número de líneas de una trama es $18432/50$ o lo que es lo mismo, 368. Puesto que cada carácter en la pantalla de 14 líneas, con 25 filas de caracteres se alcanzan las 350 líneas. El CRTC de MDA siempre utiliza 16 líneas para el retrazo vertical, esto permite 2 líneas de margen vertical: $368-(350+16)$.

La programación de CRTC sigue estos cálculos. El peso de cada carácter visualizado se especifica en el valor de registro de línea de barrido máximo del CRTC. Como los caracteres tienen una altura de 14 líneas de barrido, el valor de línea de barrido máximo es de 13 (0Dh). Considerando los valores total vertical (25 filas de caracteres) y ajuste total vertical (2 líneas de barrido) se obtiene el número total de líneas de barrido en una trama.

El número de filas de caracteres visualizado (25) está indicado en el registro de visualización vertical. La posición en la trama donde comienza el retrazo vertical (25) se especifica en el valor del registro de posición de sincronismo vertical.

5.6.1.2 Registro de estado del CRT

Existe un registro de estado del CRT (CRT status [Tabla 5.8]) de sólo lectura configurado en el puerto E/S 3BAh en el caso del MDA y el 3DAh en los adaptadores Hercules, CGA, MCGA. En EGA y VGA este registro puede estar configurado en diferentes puertos, según tengan configuraciones monocromas o de color. En el primer caso, está en el puerto E/S 3BAh, y en el segundo, en el 3DAh. Generalmente de los 8 bits que hay en este registro, dos reflejan el estado actual de las señales de barrido horizontal y vertical generados por el CRTC. Estos bits de estado se pueden utilizar para sincronizar las actualizaciones del buffer de vídeo con el ciclo de refresco de pantalla, minimizando así las interferencias con la imagen de la pantalla.

	Registro	Bit 7	Bit 3	Bit 2	Bit 1	Bit 0
MDA	3BA	0=Sinc. vertical	Controlador de vídeo		1=Trigger del lápiz óptico	1=Sinc. horizontal
HGC, HGC+, InColor	3BA		Controlador de vídeo			1=Sinc. horizontal
CGA	3DA		1=Sinc. vertical	1=Conmutación del lápiz óptico desactivada	1=Trigger del lápiz óptico	0= Activación de pantalla
EGA	3BA-3DA		1=Sinc. vertical	1=Conmutación del lápiz óptico desactivada	1=Trigger del lápiz óptico	0= Activación de pantalla
VGA	3BA-3DA		1=Sinc. vertical			0= Activación de pantalla
MCGA	3DA		1=Sinc. vertical *			0= Activación de pantalla

* 0= Sincronismo vertical en modo color de 640x480

Tabla 5.8 Asignaciones de bits para el registro de estado del CRTC

5.6.1.3 Control del modo de vídeo hardware

En general, para establecer un modo de vídeo en el subsistema del PC, se requiere, además de especificar los parámetros del CRTC, una programación del modo específico. Por ejemplo, se debe activar el generador de caracteres alfanuméricos para los modos alfanuméricos y desactivarlo en los modos gráficos. Así mismo, el reloj de caracteres interno del subsistema, que determina el número de pixels generados para cada código de carácter alfanumérico leído desde el buffer de vídeo, puede correr a diferentes velocidades según los distintos modos.

MDA

El registro de control de modo es un registro de sólo escritura configurado en el puerto 3B8h (Tabla 5.9). Sólo 3 de los 8 bits son usados. El bit 0 es puesto a 1 cuando se enciende el ordenador y debe estar siempre con ese valor. El bit 3, cuando está a 1, activa el refresco de vídeo, y cuando está a 0, borra la pantalla. El bit 5 es el bit de activación del parpadeo, que controla el posible parpadeo de los caracteres.

Bit	Asignación
0	1= Adaptador activo (siempre a 1)
1	(No usado, siempre a 0)
2	(No usado, siempre a 0)
3	1=Vídeo activo
4	0=Vídeo desactivado (pantalla en blanco)
5	1=Atributo de parpadeo activado 0=Atributo de parpadeo desactivado
6	(No usado, siempre a 0)
7	(No usado, siempre a 0)

Tabla 5.9 Asignación de bits del registro de control para MDA (3B8h)

Bit	Asignación
0	1= Modos alfanuméricos de 80 caracteres 0= Modos alfanuméricos de 40 caracteres
1	1= Modo gráfico de 320 puntos de ancho 0= Otros modos de vídeo
2	1= Burst de color desactivado (sólo en CGA) 1= Color de imagen desde el registro 7 del DAC de vídeo (sólo en MCA) 0= Burst de color activado (sólo en CGA) 0= Color de imagen desde el registro DAC de vídeo especificado en los bits 0-3 del registro de paleta (3D9h) (sólo en MCGA)
3	1= Vídeo activo 0= Vídeo desactivado (pantalla en blanco)
4	1= Modo gráfico de 640 puntos de ancho 0= Otros modos de vídeo
5	1= Atributo de parpadeo activado 0= Atributo de parpadeo desactivado
6	(No usado, siempre a cero)
7	(No usado, siempre a cero)

Tabla 5.10 Asignación de bits del registro de control para CGA y MCGA (3D8h)

CGA y MCGA

El registro de control de modo de CGA y MCGA se encuentra en 3D8h (Tabla 5.10). Los cinco bits de orden más bajo controlan los tiempos de barrido interno para los modos de vídeo seleccionados, mientras que el bit 5 es un bit de activación de parpadeo.

MCGA tiene dos registros de control de modo no implementados en CGA. El registro de control de modo del controlador de memoria (3D4h / 3D5h). Un registro de control de modo extendido configurado en el puerto E/S 3DDh. Este registro sólo se utiliza durante el arranque en frío del ordenador y no tiene un uso práctico en programas de aplicación.

Número de modo del BIOS	Descripción	Valor del registro de control de modo
0	Alfanumérico 40 X 25 (<i>Burst</i> de color desactivado)	00101100b (2CH)
1	Alfanumérico 40 X 25	00101000b (28H)
2	Alfanumérico 80 X 25	00101101b (2DH)
3	Alfanumérico 80 X 25 (<i>Burst</i> de color desactivado)	00101001b (29H)
4	Gráfico 320 X 200	00101010b (2AH)
5	Gráfico 320 X 200 (<i>Burst</i> de color desactivado)	00101110b (2EH)
6	Gráfico 640 X 200	00011100b (1CH)
7	Alfanumérico 80X25 (sólo en MDA)	00101001b (29H)
11 H	Gráfico 640 X 480 (sólo en MCGA)	00011000b (18H)

Tabla 5.11 Opciones del registro de control del modo de MDA, CGA y MCGA

Bits	Asignación
0	1= Modo 256 colores 320 X 400 0= (Otros modos)
1	Modo 2 colores 640 X 480 0= (Otros modos)
2	(Reservado)
3	1= Parámetros de barrido horizontal calculados para el modo de vídeo 0= Parámetros de barrido horizontal como se especifica en los registros 00-03H
4	1= Activa reloj de puntos (siempre a 1)
5	(Reservado)
6	1= Inverso del bit 8 del registro de Visualización vertical (06H)
7	0= Permite la actualización de los registros 00-07H

Tabla 5.12 Asignación de bits del registro de control de modo del controlador de memoria de MCGA

EGA y VGA

Cuando se establece un modo de vídeo de EGA y VGA, se puede controlar la temporización interna y el direccionamiento de los diferentes componente de vídeo del subsistema; esto incluye el secuenciador, el controlador de gráficos y el controlador de atributos, cada uno de los cuales tiene varios registros de control. También hay un registro de salida misceláneo que controla el puerto E/S y el direccionamiento del buffer de vídeo y selecciona la frecuencia del reloj de barrido.

5.6.1.3 Secuenciador

Genera la temporización interna para el direccionamiento de la RAM de vídeo. Tiene cinco registros de datos programables configurados en los puertos 3C4h y 3C5h de manera análoga al mapa de registros de CRTC (Tabla 5.15).

Bit	Asignación
0	(No usado)
1	1= Modo gráfico 720 X 348 0= Modo alfanumérico 80 X 25
2	(No usado, siempre a 0)
3	1= Vídeo activo 0= Vídeo desactivado (pantalla en blanco)
4	(No usado, siempre a 0)
5	1= Atributo de parpadeo activado 0= Atributo de parpadeo desactivado
6	(No usado, siempre a 0)
7	1= <i>Buffer</i> de modo gráfico visualizado desde B800:0000 (página de vídeo 1) 0= <i>Buffer</i> de modo gráfico visualizado desde B800:0000 (página de vídeo 0)

Tabla 5.13 Asignación de bits del registro de control de modo Hércules (3D8h). Este registro es el mismo para HGC, HGC+ y tarjeta InColor

Bit	Asignación
0	1= Permite modo gráfico 0= Evita modo gráfico
1	1= Activa los 32 KB altos del <i>buffer</i> de vídeo de modo gráfico en B800:0000 0= Desactiva los 32 KB altos del <i>buffer</i> de modo gráfico
2-7	(No usados)

Tabla 5.14 Asignación de bits del registro de conmutación de configuración Hércules (3BFh). Este registro es el mismo para HGC, HGC+ y tarjeta InColor

Registro	Nombre	Nombre original
0	<i>Reset</i>	<i>Reset</i>
1	Modo reloj	<i>Clocking Mode</i>
2	Máscara de mapa	<i>Map Mask</i>
3	Selección de mapa de caracteres	<i>Character Map Select</i>
4	Modo de memoria	<i>Memory Mode</i>

Tabla 5.15 Registros del secuenciador de EGA y VGA (3C4h y 3C5h)

5.6.1.4 Controlador de gráficos

El controlador de gráficos media tanto entre el flujo de datos del buffer de vídeo y la CPU como desde el buffer de vídeo al controlador de atributos. El controlador de gráficos tiene 9 registros de datos, además de un registro de direcciones, este último configurado en el puerto 3CEh y el registro de datos, en el 3CFh (Tabla 5.16).

Registro	Nombre	Nombre original
0	<i>Set/ Reset</i>	<i>Set/ Reset</i>
1	Activación de <i>Set/ Reset</i>	<i>Enable Set/ Reset</i>
2	Comparar color	
3	Rotar dato/ Seleccionar función	<i>Data Rotate/ Function Select</i>
4	Selección del mapa de lectura	<i>Read Map Select</i>
5	Modo gráfico	<i>Graphics Mode</i>
6	Misceláneo	<i>Miscellaneous</i>
7	<i>Color Don't Care</i>	<i>Color Don't Care</i>
8	Máscara de bit	<i>Bit Mask</i>

Tabla 5.16 Registros del controlador de gráficos de EGA y VGA (3CBh y 3CFH)

5.6.1.5 Controlador de atributos

Soporta una paleta de 16 colores tanto en EGA como en VGA. También controla el color de la pantalla durante los intervalos de imagen. El registro de direcciones del controlador de atributos y los 21 registros de datos están configurados en el puerto E/S 3C0h. Dependiendo del valor de un biestable interno del controlador de atributos, los valores escritos en el puerto 3C0h son almacenados en el registro de direcciones o en un registro de datos. Para establecer el biestable, se realiza una lectura E/S (IN AL, DX) del registro de estado del CRT (3BAh para monocromos, 3DAh para color) (Tabla 5.17).

Registro	Función
0-0Fh	Paleta
10h	Control de modo de atributo
11h	Color del margen
12h	Activación del plano de color
13h	Panorámica de pixel horizontal
14h	Selección de color (sólo en VGA)

Tabla 5.17 Registros del controlador de atributos de EGA y VGA (3C0h y 3C1h)

5.6.2 MODOS ALFANUMÉRICOS

A excepción del MDA (que no tiene modo gráfico), el resto de los sistemas de vídeo del PC y PS/2 se pueden programar para presentar los caracteres tanto en modo gráfico como alfanuméricos.

5.6.2.1 Representación de datos alfanuméricos

Cada carácter se representa mediante una sencilla estructura de datos de dos bytes (Fig. 5.19). Los cuales se almacenan en el buffer en una secuencia lineal que se dirige hacia la parte derecha e interior de la pantalla.

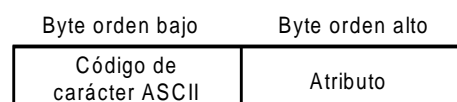


Fig. 5.19 Codificación de un carácter. Código ASCII y atributo de pantalla

El generador de caracteres hardware convierte cada código de carácter en el patrón adecuado de puntos de pantalla. Al mismo tiempo, la circuitería del decodificador de atributos genera los atributos adecuados de cada carácter: color, intensidad (brillo, parpadeo,...) puesto que cada código de carácter del buffer de vídeo está acompañado por un byte de atributos, se pueden controlar independientemente los atributos de visualización para cada carácter en la pantalla.

El generador de caracteres hardware visualiza cada carácter alfanumérico en una matriz rectangular de pixels. Dentro de esa matriz, el carácter está formado por un conjunto de pixels de fondo. Los colores de los pixels de fondo e imagen del carácter se especifican con los nibbles alto y bajo del correspondiente byte de atributos.

5.6.2.2 Atributos

Un byte de atributo puede interpretarse de varias formas. En general se forma el byte de atributos con 2 nibbles de 4 bits; el byte de orden más bajo (desde el bit 0 hasta el 3) determina los atributos de imagen del carácter, esto es el color e intensidad del carácter. El nibble de orden más alto indica los atributos de fondo de carácter, aunque el bit 7 también puede controlar el parpadeo en situaciones determinadas.

MDA

Aunque se pueden especificar cualquiera de los 16 (2^4) atributos tanto para los atributos de fondo como para los de imagen, MDA sólo reconoce determinadas combinaciones. Sin embargo, puede generar una variedad útil de atributos de carácter combinando adecuadamente la intensidad, parpadeo y subrayado.

El bit 7 del byte de atributo puede servir como controlador del parpadeo definiéndolo a 1 si el bit de registro de control de modo CRT está puesto a 1. Si este bit de activación de parpadeo es 0, el bit 7 se utilizará como un bit más de intensidad.

CGA

La CGA utiliza el mismo esquema de atributos para imagen y fondo de la MDA. Sin embargo la circuitería de decodificación de atributos de la CGA reconoce cualquiera de las 16 combinaciones posibles de los cuatro bits de cada nibble del byte de atributos.

Los colores disponibles son combinaciones sencillas de los colores primarios rojo (R), verde (G), azul (B) (Fig. 5.20).

El bit de orden más alto (bit 7) de cada byte de atributos controla tanto la intensidad de fondo como el parpadeo, dependiendo del estado de un bit dentro del registro de control.

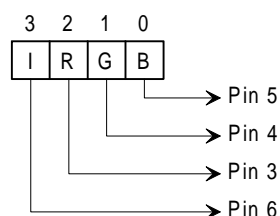


Fig. 5.20 Asignación de pines en el conector.

EGA

En los modos alfanuméricos de 16 colores, EGA utiliza el mismo formato de byte de atributos de CGA. Sin embargo, los valores de los 4 bits de fondo e imagen no se corresponden

directamente con los colores presentados en pantalla. En vez de esto, cada valor de 4 bits está enmascarado con los 4 bits de orden más bajo del registro de activación del plano de color del controlador de atributos, el valor de los 4 bits restantes designan uno de los 16 registros de paleta de EGA (Fig. 5.21). Cada bit de los 6 bits de color, que se encuentran en el registro de paleta designado, se corresponde con una de las seis señales RGB, 3 que controlan el monitor.

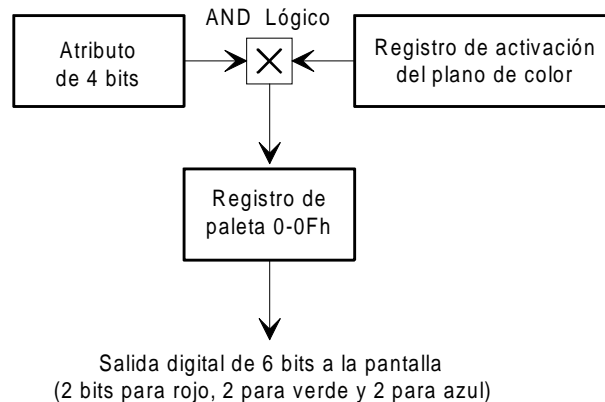


Fig. 5.21 Atributos y colores de EGA

Un monitor color compatible con EGA se controla con seis señales de color: tres primarios (intensidad alta) y tres secundarios (intensidad baja). Todas las 64 combinaciones de estas seis señales se presentan como intensidades y/o colores diferentes (Fig. 5.22).

El método que utiliza EGA para generar colores indirectamente a través de los registros de paleta es más complejo que el esquema directo de CGA, pero mucho más flexible. Se puede seleccionar y además, producir cambios globales en el color actualizando los valores de un registro de paleta determinado.

El bit 3 del registro de control de modo del controlador de atributos de EGA (registro 10h en el puerto E/S 3C0h) es el bit de activación de parpadeo. Poniendo este bit a 1 se activa el parpadeo, así que sólo los 3 bits de orden más bajo del nibble de fondo designan los registros de paleta. De este modo, cuando el parpadeo está activado sólo podemos referenciar los ocho primeros registros de paleta para seleccionar el color de un carácter.

VGA

En general, VGA emula la decodificación de atributos alfanuméricos de EGA. Sin embargo, VGA tiene tanto un DAC de vídeo como un juego de 16 registros de paleta del controlador de atributos. Cada valor de los registros de paleta selecciona uno de los 256 registros de color del DAC de vídeo. El valor del registro DAC de vídeo seleccionado determina el color que se visualiza en la pantalla.

Dependiendo del valor del bit 7 del registro de control de modo del controlador de atributos se pueden utilizar los valores de los registros de paleta para seleccionar el registro de color del DAC de vídeo de dos formas distintas:

- Cuando está a 0 (Fig. 5.23), el controlador de atributos combina el valor del registro de paleta de 6 bits (con los que selecciona un registro dentro de un grupo de 64) con los bits 2 y 3 de su registro de selección de color 14h (con los que se selecciona uno de los 4 grupos de 64 registros de color del DAC), para producir un valor de 8 bits.

- Cuando está a 1 (Fig. 5.24), sólo los cuatro bits menos significativos de cada registro de paleta son tenidos en cuenta (seleccionan uno de los 16 registros del grupo) y los cuatro bits restantes se

obtienen a partir de los bits del registro de selección de color (con lo que se selecciona uno de los 16 grupos de los registros de color del DAC).

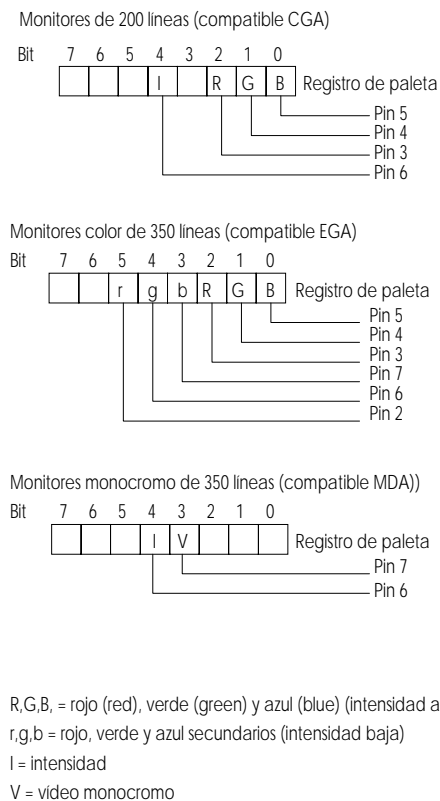


Fig. 5.22 Valores del registro de paleta de EGA y señales de control del monitor

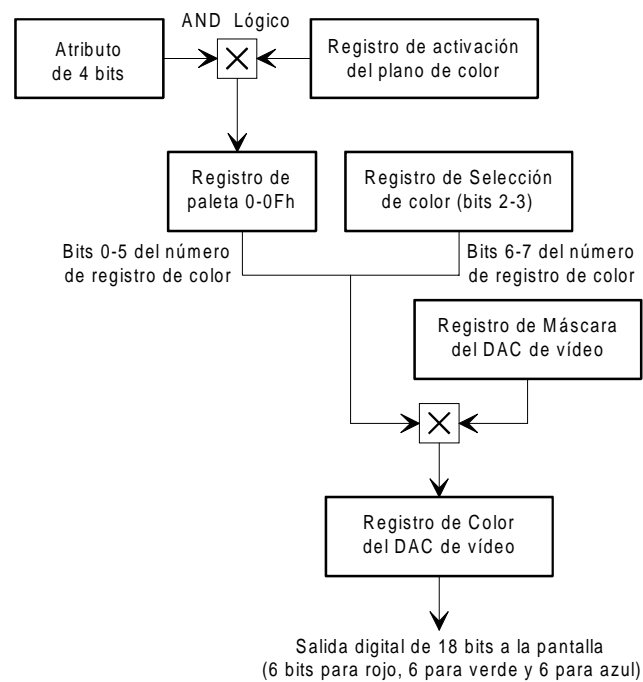


Fig. 5.23 Atributos y colores de VGA
 (cuando el bit 7 del registro de control de modo del controlador de atributos está a 0)

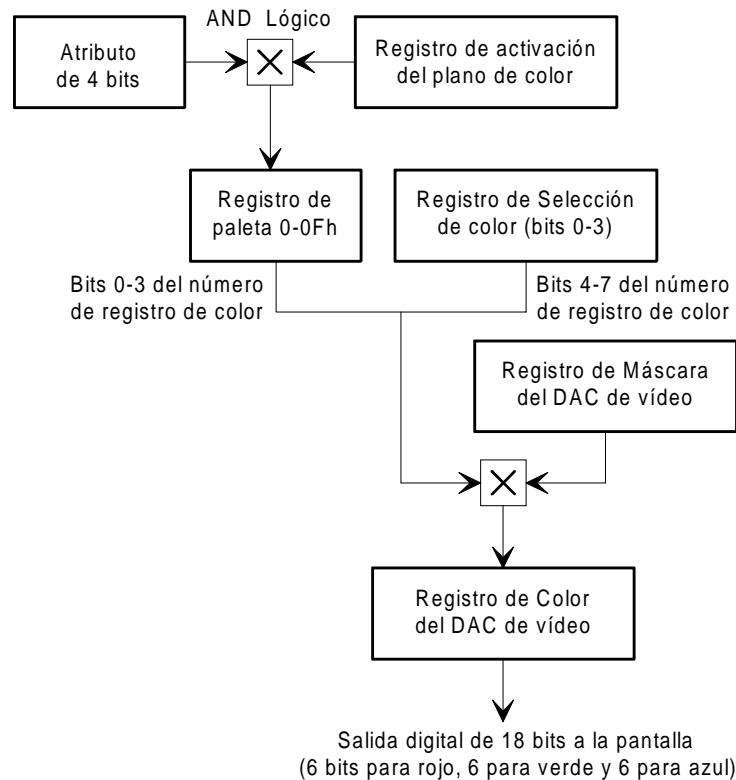


Fig. 24 Atributos y colores de VGA
(cuando el bit 7 del registro de control de modo del controlador de atributos está a 0)

5.6.3 MODOS GRÁFICOS

En los modos gráficos se puede manejar el color de cada pixel individual de la pantalla, por lo que el número de datos que se maneja es mayor. Esto se reflejará en menores prestaciones, en el caso de que el sistema no sea lo suficientemente potente.

Los subsistemas de vídeo almacenan los valores de los pixels como un grupo de bits que los representan, de modo que dicho valor determina directa o indirectamente el color del pixel correspondiente. El formato del mapa de pixels o mapa de bits dentro del buffer de vídeo depende del número de bits requeridos para representar cada píxel, así como de la arquitectura de la RAM de vídeo (su capacidad). Obviamente, el número de colores que podemos visualizar al mismo tiempo en un determinado modo gráfico está restringido por el número de bits utilizados para representar cada pixel.

Cuando los valores de pixel ocupan menos de ocho bits, los pixels se mapean en campos de bits de izquierda a derecha dentro de los bytes. Esto es válido para todos los subsistemas de vídeo del PC y PS/2.

CGA

En CGA cada pixel se representa mediante 2 bits si estamos en el modo 320x200 con 4 colores y mediante un bit si estamos en el modo 640x200 con 2 colores.

El valor de un pixel se mapea en dos mitades entrelazadas de 16 Kbytes del buffer. El valor de pixel para las 100 líneas pares comienza en B8000:0000, y para las impares, en B800:0000 (Fig. 5.25).

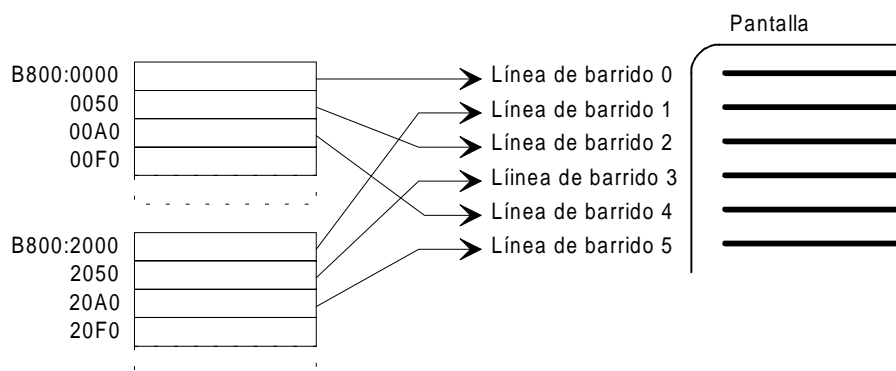


Fig. 5.25 Entrelazado del buffer de vídeo en los modos gráficos de CGA

HERCULES (HGC)

La representación de pixels en los modos gráficos 720x348 de HGC y HGC+ es similar al modo 640x200 de CGA. Cada pixel está representado por un bit, que indica encendido o apagado (blanco y negro).

Sin embargo, las 348 líneas de 90 bytes (720 columnas / 8 bits por byte) cada una, se entrelazan de manera distinta: usa cuatro áreas del buffer de vídeo, cada una de ellas conteniendo 87 (348 líneas / 4 áreas) líneas (Fig. 5.26)

EGA

Cuando se configura una EGA para que emule un modo gráfico CGA, los pixels se mapean en el buffer de vídeo igual que en la CGA. Sin embargo, en los modos gráficos propios de EGA (modo de 16 colores y 200 líneas y todos los modos de 350 líneas), los pixels se mapean siempre ocho por byte (Fig. 5.27).

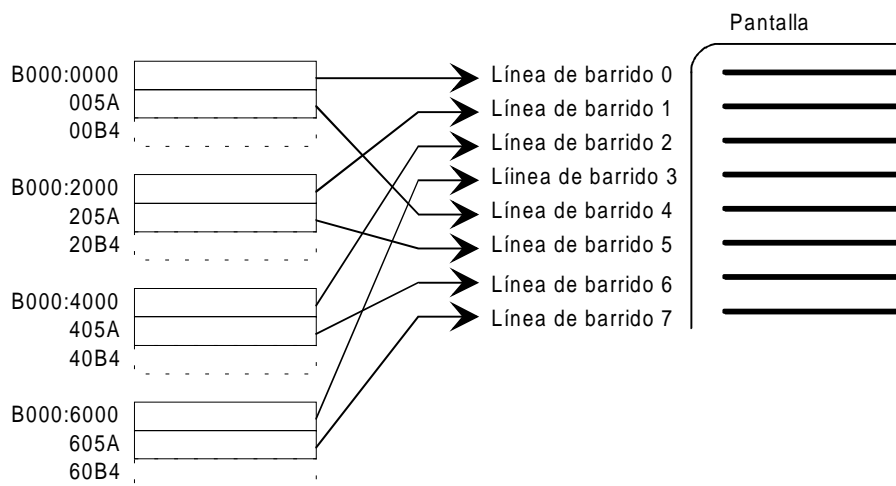


Fig. 5.26 Entrelazado del buffer de vídeo en el modo gráfico de Hércules

Esta configuración está dictada por la arquitectura del buffer de vídeo de EGA. Los 256 Kb del buffer se dividen en cuatro mapas o bancos paralelos de RAM, de 64 Kb cada uno. Estos mapas son paralelos en el sentido de que ocupan el mismo rango de direcciones en el banco de direccionamiento de la CPU; el secuenciador permite accesos a los mapas de forma individual o bien paralela.

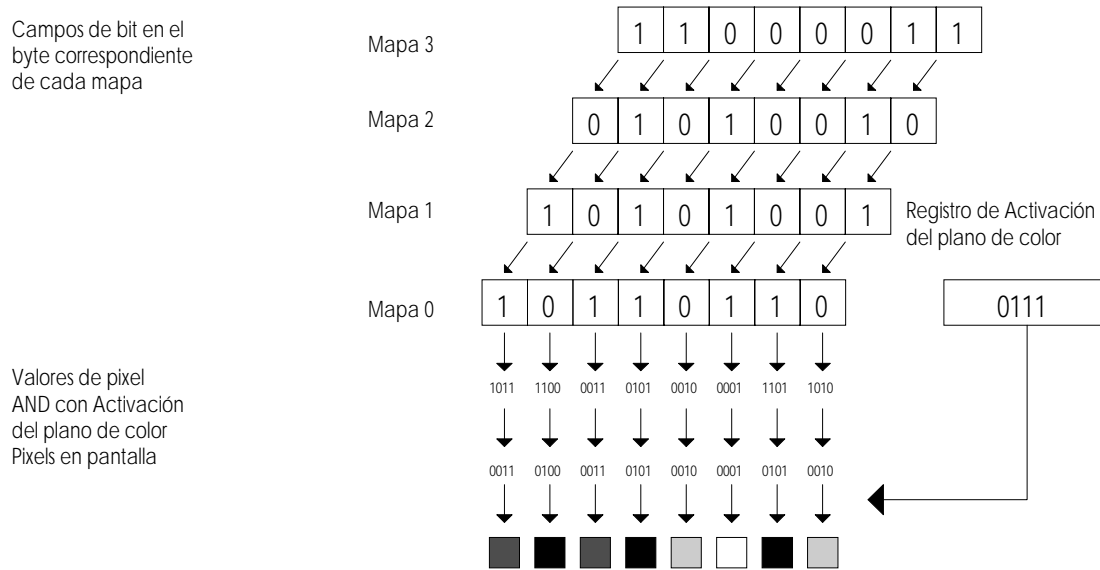


Fig. 5.27 Mapeo de pixels en los modos gráficos originales (propios) de EGA

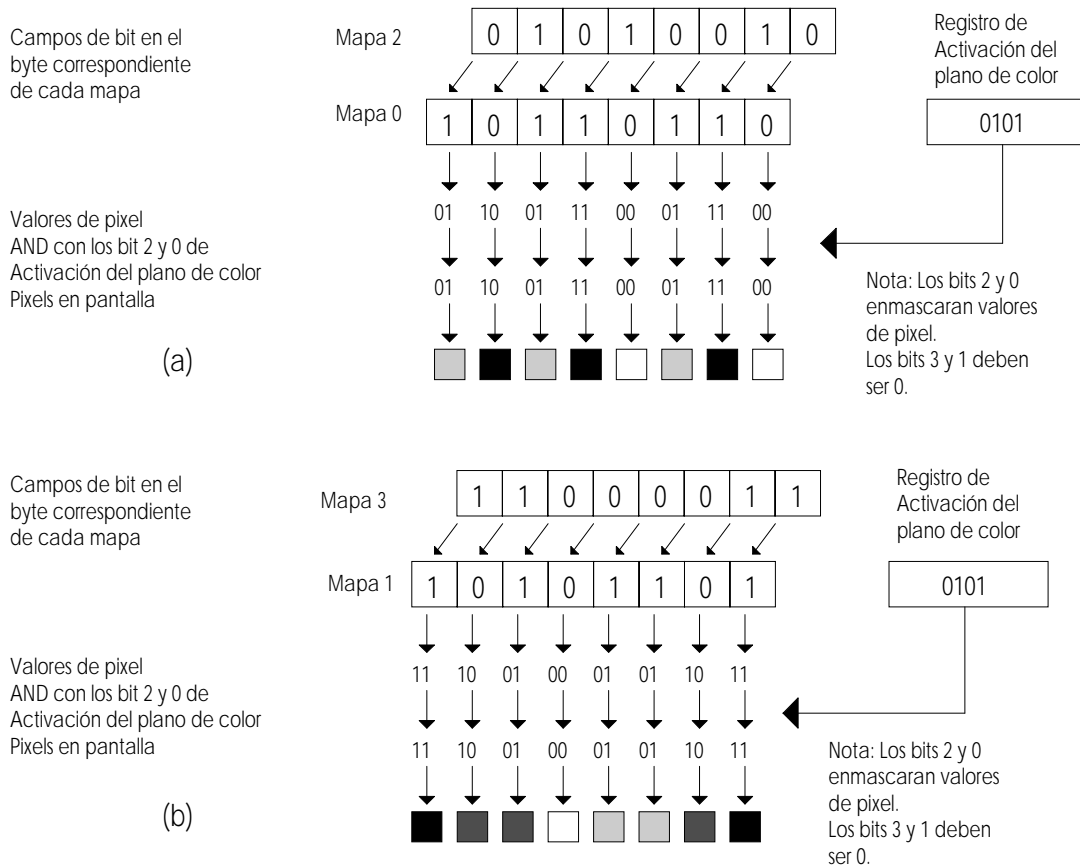


Fig. 5.28 Buffer de vídeo en los modos gráficos de 350 líneas (EGA con 64 Kb de RAM)

Los valores de pixel en las direcciones pares son almacenados en los mapas 0 y 2 (figura 4.5a); los pixels en las direcciones impares son almacenados en los mapas 1 y 3 (figura 4.5b)

Un valor de pixel viene determinado por la concatenación de los bits (determinados por un cierto desplazamiento) de los cuatro mapas. Por ello, los mapas son también denominados planos de bits (Fig. 5.28)

MCGA y VGA

Los subsistemas de vídeo del PS/2 soportan tres modos gráficos nuevos, es decir, que no estaban en los adaptadores de vídeo anteriores: el modo 640x480 con 2 colores (MCGA y VGA) y el modo 640x480 con 16 colores (sólo VGA). Ambos utilizan un mapa de pixel lineal que comienza en A000:0000. También se utiliza un mapa de pixel lineal parecido en el modo de 256 colores 320x200 (MCGA y VGA), con una diferencia importante: cada byte en el buffer de vídeo representa un pixel, por lo que cada pixel puede tener hasta 256 ($=2^8$) colores diferentes.

Periféricos de salida

6.1 INTRODUCCIÓN

Las impresoras, con las pantallas y teclados, completan el trío de periféricos de I/O (entrada/salida) estándar, que forman parte de todos los sistemas computadores comerciales y científicos. El propósito del ordenador es almacenar, procesar y proporcionar información que necesitaremos más adelante. Aunque la mayoría de las veces, esta información estará guardada en un medio de almacenamiento masivo, normalmente necesitamos tener esta información en el papel, y es por ello que la impresora es un componente básico del sistema. Muchas impresoras modernas pueden manejar material gráfico bastante bien, pero donde el principal propósito del sistema es producir gráficos de líneas, el plotter es la mejor elección. Los plotters son muy apreciados en entornos de ingeniería y arquitectura y en general en tareas de diseño asistido por ordenador (CAD: Computer Assisted Design).

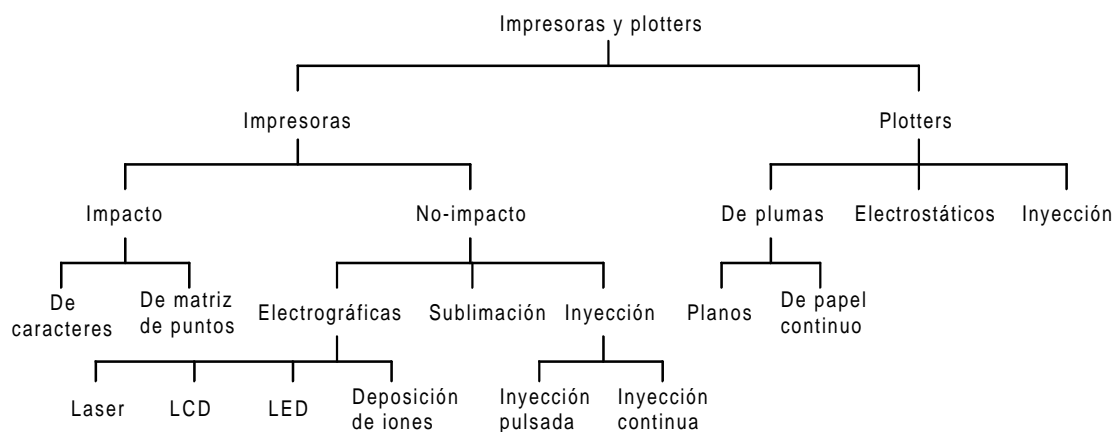


Fig. 6.1 Principales tipos de impresoras y plotters

La distinción entre impresoras y plotters, es que las impresoras construyen las imágenes o texto de cada página sistemáticamente, línea a línea, usando un conjunto límite de elementos discretos (caracteres, símbolos o puntos) mientras que los plotters crean la imagen dibujando líneas, moviendo un lápiz o pluma sobre el papel como se hace cuando se dibuja en un papel.

Dentro de esta amplia distinción existen muchos tipos de impresoras y varios tipos de plotters. Nosotros consideraremos sólo los más importantes de ellos.

Las impresoras pueden clasificarse de diferentes formas. Una posible clasificación distingue entre impresoras de páginas, líneas y caracteres según sea la operación simple a imprimir. Sin embargo, las impresoras de carácter, son controladas normalmente para leer una línea completa en un buffer interno antes de imprimirla, por lo que la posibilidad de escribir un carácter cada vez no está disponible para el usuario.

Las impresoras pueden clasificarse también en 'Solid Font Printers' (o de caracteres) y 'Matrix Printers' (o de matriz de puntos). Las impresoras de caracteres (o 'Solid Font') tienen un conjunto de caracteres predefinidos, como las máquinas de escribir. Las impresoras de matriz de puntos construyen lo que van a imprimir a partir de un array de puntos, con lo que pueden imprimir cualquier imagen. Una tercera división es entre las impresoras de Impacto, las cuales trabajan como las máquinas de escribir, y las de No Impacto, que utilizan otros métodos más silenciosos (Fig. 6.1).

6.2 IMPRESORAS DE IMPACTO

6.2.1 Máquinas de escribir y teletipos

Antes de que las pantallas de video (VDU o 'Vídeo Device Unit') se volvieran de uso común en 1960, la manera más usual para que el operador controlara y se comunicara con la computadora era en general un dispositivo periférico, que combinaba un teclado con una simple impresora. Dos de los dispositivos ya existentes fueron adoptados para el uso con la computadora. Uno de estos dispositivos fue la Teleimpresora o 'Teletipo'. El otro dispositivo fue la máquina de escribir eléctrica. Esta fue más popular que el teletipo, pero precisó de alguna modificación para hacerla compatible con la computadora.

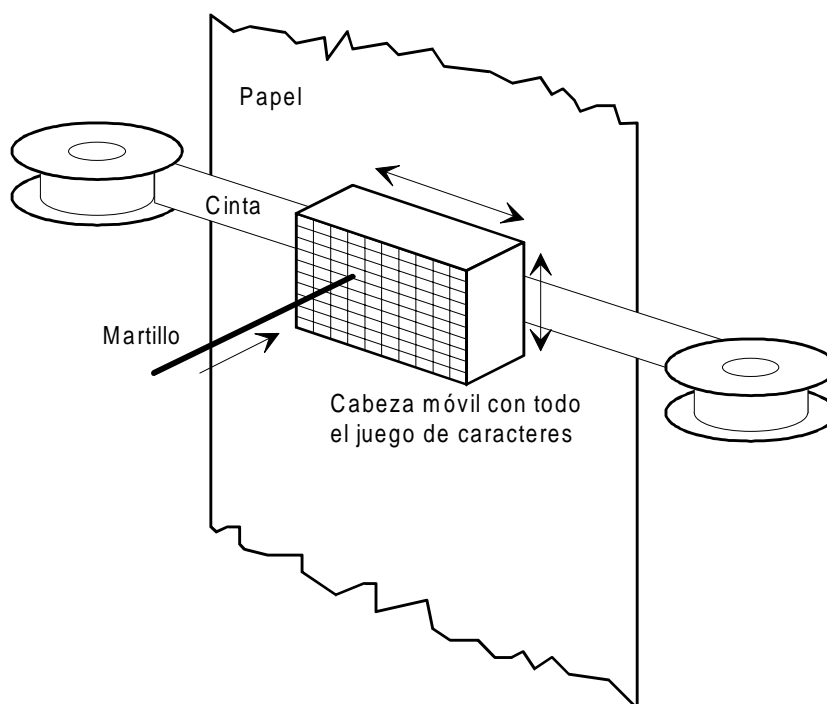


Fig. 6.2 Mecanismo de impresión de las teleimpresoras

La Teleimpresora se fabricó originalmente para el enlace en las comunicaciones de telegramas y Telex (Fig. 6.2). Presionando una de las teclas del teclado se envía un código que representa a un carácter de otra teleimpresora en el otro extremo del enlace. Aquí, un martillo golpea sobre el papel interponiendo entre ellos una cinta con tinta, la cual hace que el carácter se imprima. Cada impresora tiene un conjunto de 64 caracteres en una matriz rectangular. El código del carácter se identifica con una posición en la caja, y la caja es movida hasta enfrenar el martillo al carácter. La cabeza de impresión completa es entonces movida (incluyendo la matriz de caracteres y el mecanismo de martilleo) a través del papel, a la posición donde aparecerá el próximo carácter. La cinta con tinta es enrollada también en el ancho de un carácter, preparándose para imprimir el siguiente carácter sobre el trozo de cinta adyacente. Cuando el carácter que se envía es un retorno de carro, la cabeza de impresión es retornada a la parte izquierda del papel y éste se mueve hacia arriba una línea, preparándose para comenzar en la siguiente. Usualmente (pero opcionalmente), cada uno de los caracteres se imprime también en la teleimpresora que envía el mensaje, con lo que tendremos una copia del mensaje en cada uno de los terminales del enlace. Era común que los mensajes salientes se escribieran en rojo, y los entrantes en negro. Para permitir esto, se utiliza una cinta de dos colores. La mitad superior de la cinta será roja y la inferior será negra, moviéndose la cinta de tal forma, que enfrentaremos el color requerido a la zona de impresión. Cada cinta es utilizada muchas veces, ya que es reversible. Cuando la cinta llega al final, el carrete que aportaba cinta ya no puede continuar haciéndolo, de forma que la cinta se tensa más de lo habitual y esto activa un mecanismo que invierte el sentido de giro de los carretes. De esta forma el carrete que hasta ahora ha estado enrollando la cinta pasará a proporcionarla y ésta se irá enrollando en el carrete en el que estaba inicialmente.

Los caracteres de la matriz pueden ser removibles, por lo que cualquier carácter puede estar en cualquier posición de la caja, aunque por supuesto, ello podría ser un inconveniente si los caracteres no se corresponden con la letra del teclado. Puesto que el número de posiciones es pequeño, sólo algunas figuras, y en su caso, letras mayúsculas pueden ser incluidas. Unos pocos caracteres se usan como caracteres de control (tales como el Retorno de Carro), por lo que, de hecho, pueden ser impresos, menos de 64 caracteres.

Un diseño alternativo, consiste en que el array de elementos de impresión está curvado en forma de cilindro. Ahora sólo uno de los movimientos necesarios para seleccionar el carácter es lineal, mientras que el otro es un giro. En este caso, todos los caracteres están puestos en la cara de un cilindro sólido, con lo que el alfabeto es fijo. Un principio similar fue utilizado en las máquinas de escribir de bola en las que los caracteres se disponían sobre una esfera. El aspecto de esta esfera era similar al de una pelota de golf por lo que a estas máquinas se las conocía como máquinas 'Golf ball'.

Las máquinas teletipo son lentas (alrededor de 7 caracteres por segundo) y ruidosas, además de tener un conjunto de caracteres restringido, pero fueron atractivas para los primeros diseñadores de computadoras, puesto que podían usarse sin ninguna modificación. Fue sólo necesario fijar la máquina al ordenador con un interfaz, que emulara a otra máquina teletipo. Por lo tanto, el teletipo se convirtió en la consola estándar para las primeras computadoras, antes de que se introdujera el VDU para reemplazarla.

Las máquinas de escribir eléctricas ya estaban disponibles cuando se iban introduciendo los ordenadores, y algunos de éstos se adaptaron para que se usasen como terminales del operador del ordenador en lugar de los teletipos. Eran más caras que los teletipos, pero daban una mejor calidad de impresión y un rango de caracteres más amplio, incluyendo las letras en minúscula. Algunas estaban disponibles para manejar papel más ancho, mientras que los teletipos estaban limitados a unas 8 pulgadas. Las primeras versiones, tales como los 'Flexowriter' disponían de un muelle para conducir cada tecla a su sitio cuando ésta era pulsada por el dedo. Más tarde IBM introdujo las máquinas de escribir 'Golf ball', en las que el enlace entre las teclas y el mecanismo de impresión era eléctrico en lugar de mecánico, y también el acarreo del papel permanecía estacionario,

mientras que la cabeza se movía a través de la página. Esto se adaptó mucho más fácilmente para el uso como periférico de un ordenador. También era mucho más rápido, alrededor de 15 caracteres por segundo, que las primeras adaptaciones de máquinas de escribir o teletipos, aunque bastante caro. Los elementos a imprimir estaban sobre una cabeza de superficie esférica que se rotaba e inclinaba para llevar el carácter requerido a la posición de impresión.

Los teletipos y máquinas de escribir se adecuaron para los computadores científicos, pero cuando los ordenadores comenzaron a usarse comercialmente, el volumen de las impresoras resultó demasiado grande, y se hicieron precisas impresoras mucho más rápidas. De nuevo, una máquina conveniente existía ya, esta vez fue el ahora olvidado campo de las tarjetas perforadoras. Esta máquina, llamada 'tabulador', leía los datos de una tarjeta perforada; cada tarjeta almacenaba sobre 80 caracteres alfanuméricos y se correspondía con una línea de impresión. El método de impresión fue el de conducir de nuevo el martillo al 'Slog' como en los teletipos, pero en este caso era un conjunto de 80 'Slogs' y su mecanismo de martillo. De esta forma, una línea completa era impresa a la vez, usando un cinturón entintado del ancho del papel, más que una cinta estrecha. Para permitir el espaciado normal entre caracteres, cada conjunto de 'Slogs' se arregló como una simple columna más que como un array rectangular. Por supuesto, no tenía teclado y no había dificultad para modificar esta máquina para imprimir datos alimentados desde el ordenador principal o desde las tarjetas perforadas. Esta máquina imprimía alrededor de dos líneas por segundo.

Estos tres tipos de máquinas, todas impresoras de impacto, fueron el sostén principal de los primeros ordenadores. Sin embargo, todas tenían unas limitaciones, especialmente de velocidad, y como los ordenadores industriales crecieron, se hizo necesario diseñar dispositivos según las necesidades del ordenador. También se introdujeron los VDU, siendo muy convenientes como estación de operación, donde los mensajes eran usualmente transitorios. Por lo tanto, las impresoras que ya hemos descrito se volvieron obsoletas.

6.2.2 Impresoras de margarita

Las impresoras 'daisy-wheel' o de margarita (Fig. 6.3) trabajan de una forma similar a las máquinas de escribir 'Golf ball' y, efectivamente, ahora también hay máquinas de escribir de margarita. Reciben este nombre debido a que el relieve que producen las letras sobre la esfera recuerda los hoyuelos de las pelotas de golf. El computador imprime, y sin embargo, no tiene un teclado pegado. Se trata de una impresora de impacto puro, e imprime un carácter cada vez, aunque en la mayoría de los casos, se acepta una línea completa como entrada desde el host, antes de imprimir.

Como en los teletipos, la impresora utiliza un martillo para golpear el carácter contra la cinta entintada y el papel. Para la impresión completa, la cabeza se mueve un carácter después de la impresión del primero. Sin embargo, en este caso, los caracteres no están separados, sino formados al final de unos 'pétalos' largos y flexibles, que parten radialmente desde un eje central. Este tipo de ensamblaje, de pétalos y eje, se llama 'printwheel', o más comúnmente margarita ('daisy wheel'). La rueda se rota hasta que el carácter apropiado se enfrenta al martillo, para la impresión. La margarita se fabrica de una sola pieza, y a menudo se moldea en plástico. Por tanto, no es posible reemplazar caracteres individualmente, aunque resulta fácil reemplazar la margarita completa. Esto permite variar el tipo y tamaño, usualmente 10 ó 12 caracteres por pulgada, aunque algunas impresoras y margaritas proporcionan un espaciado proporcional. Esto también permite el uso de conjuntos de caracteres especiales, para imprimir por ejemplo símbolos matemáticos o un alfabeto extranjero.

La mayoría de las margaritas se diseñan para imprimir 96 caracteres (que es un ASCII completo menos los caracteres de control), entre los que se incluyen las letras mayúsculas,

minúsculas, y los signos de puntuación. La margarita tiene pues, 96 pétalos. Algunas impresoras usan doble margarita. Esta tiene menos pétalos, en cada uno de los cuales hay dos caracteres (como las máquinas de escribir que tienen 2 caracteres en cada pulsador). La margarita se mueve arriba y abajo para realizar la selección entre ellos. Tales impresoras de doble margarita, a menudo poseen un conjunto de 128 caracteres.

Como el mecanismo de impresión es de un carácter a la vez, se puede retroceder, con el objeto de que los caracteres subrayados se escriban con dos pulsaciones. En la mayoría de estas máquinas, también puede hacerse una impresión doble o enfatizada, retrocediendo la cabeza e imprimiendo una segunda línea de caracteres, normalmente con un pequeño desplazamiento horizontal, para dar el énfasis.

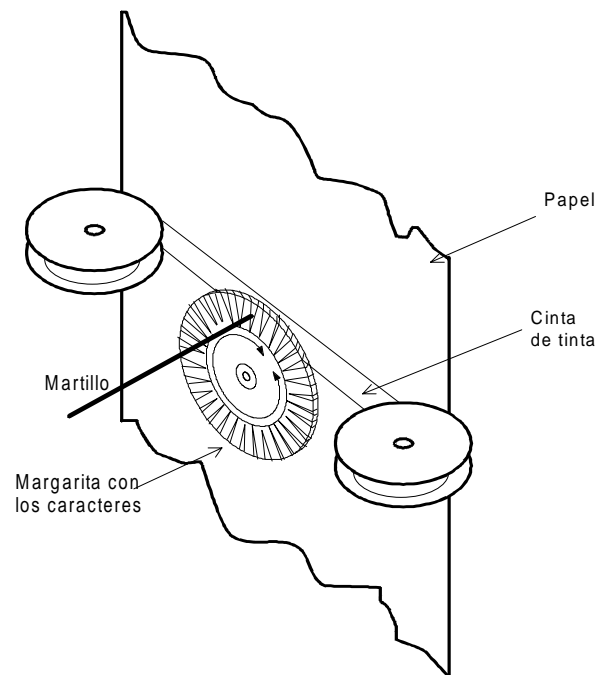


Fig. 6.3 Mecanismo de una impresora de margarita. El soporte de impresión tiene forma de margarita con un carácter en cada "pétalo"

Las impresoras de margarita usan una cinta estrecha que se mueve con cada pulsación, al igual que en las máquinas de escribir y en los teletipos. Puede ser usual que la cinta impregnada de tinta se enrolle de un carrete a otro, que es reversible de forma que la cinta se reutiliza cuando alcanza el final. Una alternativa es usar una cinta interminable que se introduce en un contenedor (sin carretes), y se saca por el final. Esto evita los mecanismos de reversión de la cinta. Además, también permite que la cinta sea girada sobre sí misma en el paso a través del contenedor, de tal forma que la golpeamos por la otra cara en la siguiente pasada. Algunas impresoras de margarita (al igual que algunas máquinas de escribir), pueden usar cintas de un sólo uso. Estas se basan en carbón, el cual se transfiere por completo al papel de un golpe, y la cinta no es reversible. Estas cintas dan una mejor calidad de impresión, pero por supuesto, su uso resulta más caro.

Las margaritas no suelen llevar caracteres extras del 'IBM Graphics Set', excepto ruedas especiales en las que algunos caracteres estándares pueden haber sido reemplazados por otros. Las impresoras de margarita, pues, tienen un restringido número de caracteres a imprimir, y no pueden manejar gráficos. Además, son más bien lentas y ruidosas. Las impresoras de matriz son mejores en ambos aspectos, aunque con peor calidad de impresión (y no pueden utilizar cintas de un sólo uso). Las impresoras láser ofrecen una buena calidad de impresión y velocidad, pero son más caras que las anteriores. Por tanto, las impresoras de margarita quedaron restringidas a algunas

aplicaciones (tales como la correspondencia comercial) donde la calidad de impresión era fundamental y la velocidad no era crítica. Actualmente han quedado desbancadas por las mejores prestaciones de las impresoras láser.

Una impresora de margarita típica imprime de 20 a 30 caracteres por segundo, aunque se ha conseguido hasta 80 caracteres por segundo. Nótese sin embargo, que "caracteres por segundo" no es siempre un valor básico para la comparación de la velocidad entre las impresoras de carácter por golpe, puesto que el tiempo que se toma para el retorno de carro y la alimentación del papel afecta al rendimiento; y ello varía según el diseño. En particular, la impresión bidireccional es posible en algunas impresoras. En este caso, líneas alternas son impresas siguiendo direcciones opuestas. Este es el motivo fundamental de que estos tipos de impresoras no impriman caracteres individuales sino sólo líneas completas.

6.2.3 Impresoras de barril

Al contrario que los teletipos y las impresoras de margarita, las impresoras de barril (Fig. 6.4) dan una línea de impresión cada vez, y están diseñadas para una mayor rapidez de impresión, como sucesoras de las máquinas tipo columna basadas en 'tabuladores tarjetas punzón'. Se popularizaron en los años 60, pero ahora están obsoletas; se sustituyeron por las de fila, de banda y de línea (impresoras de matriz), y más recientemente, por las impresoras de no impacto.

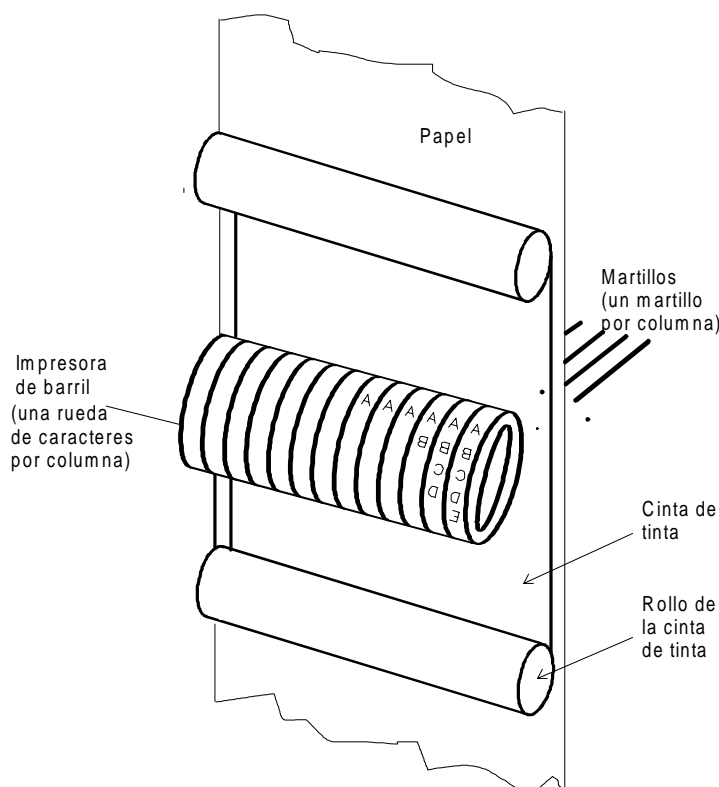


Fig. 6.4 Mecanismo de las impresoras de barril

Estas son, de nuevo, impresoras de impacto y 'Solid font'. El mecanismo es, en algunos aspectos, un cruce entre las impresoras de margarita, y las de 'Golf ball'. Los caracteres se encuentran en circunferencias, formando un cilindro, y éste gira a velocidad constante alrededor de un eje horizontal. Sin embargo, un conjunto de caracteres se mueve cada vez a su posición de impresión, y hay un conjunto separado para cada una de las posiciones de impresión en la línea (a menudo 132). Cada uno de los conjuntos puede fabricarse separadamente como una rueda de impresión ('printwheel'), pero todas las ruedas están ensambladas en un cilindro rígido que gira

como una unidad simple. Se pone una cinta del ancho del papel, entre el tambor y el papel. El martillo es movido durante un instante, cuando la letra requerida está enfrenteada a él. El impacto debe ser muy corto, o el carácter saldrá manchado. El tiempo en el que el martillo golpea al carácter es también crítico, pues de lo contrario el carácter saldrá en una posición vertical errónea; asimismo, debe ajustarse la fuerza del impacto, para que todos los caracteres tengan el mismo tono de negro. En realidad, la característica de la salida de las impresoras de barril, es una línea ondulada de impresión, con densidad variada, puesto que el ajuste correcto es muy difícil, siendo ésta una de las razones por las que esta impresora cayó en desuso.

Las primeras impresoras de barril podían imprimir alrededor de 150 líneas por minuto; diseños posteriores llegaron a las 600 líneas por minuto, lo que equivale a 10 líneas por segundo o unos 1300 caracteres por segundo.

6.2.4 Impresoras de banda de cadena y de tren

Este tipo de impresoras es similar, en principio, a las impresoras de barril, y, como ellas, tienen líneas de impresión por impacto: 'Solid font'. Sin embargo, los elementos, en vez de moverse verticalmente, se mueven horizontalmente. Para conseguir esto, tales elementos se emplazan en una banda de impresión, que es una banda o cinta de acero flexible que pasa sobre un par de poleas, una en cada extremo del papel, y localizada frente a la línea de impresión (Fig. 6.5). Ahora no es necesario todo el conjunto de caracteres en cada posición del carácter; un conjunto de caracteres lo hará todo. De este modo, cada carácter pasa a través de cada posición de impresión de la línea al mover el cinturón. De nuevo, eso sí, hay un martillo para cada posición de impresión, cada uno de los cuales se acciona cuando el carácter requerido pasa frente a él. Aún existe el problema del emborronamiento y el ajuste de los martillos para localizar el carácter correctamente pero no existen variaciones en el nivel de negro del carácter. Sin embargo, los errores de tiempo se muestran ahora como variaciones en el espaciamiento de los caracteres, en vez de en el alineamiento, por lo que resulta más estético.

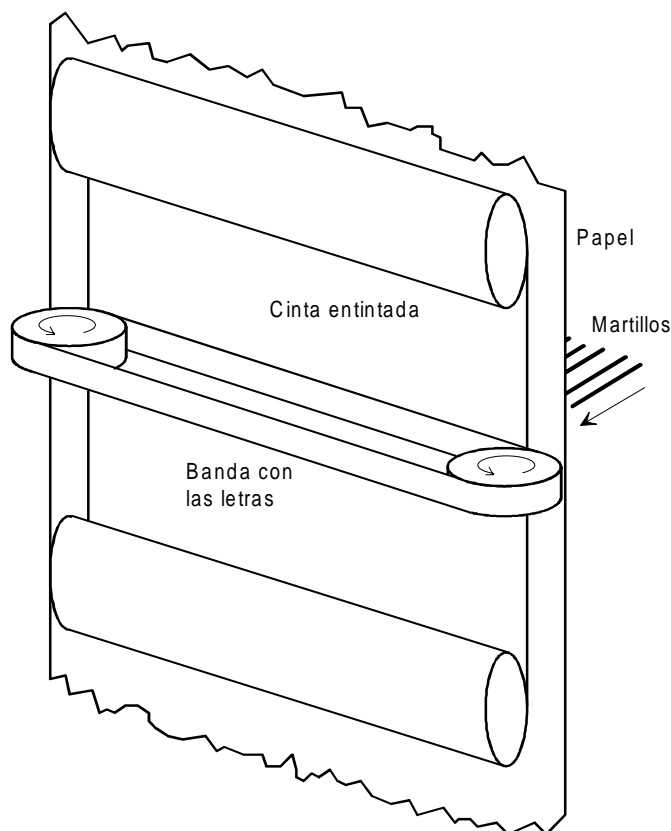


Fig. 6.5 Mecanismo de las impresoras de banda

En la práctica, y en interés del rendimiento, el conjunto de caracteres está repetido más de una vez en la banda -típicamente 4 veces-. No es necesario repetir cada carácter el mismo número de veces. Por ejemplo, puede haber en la banda 8 veces la 'e', y sólo una vez la 'ñ'. Usualmente, la banda es de metal, con los caracteres grabados sobre ella. Es imposible, por lo tanto, cambiar individualmente un carácter, pero no es demasiado difícil (aunque es trabajo del técnico) cambiar la banda por otra con un conjunto diferente de caracteres. La impresora de banda se hizo popular a finales de los años 70, y aún se usa en grandes sistemas. Las impresoras de banda pueden imprimir sobre 2500 líneas por minuto; unos pocos modelos tienen un conjunto doble de martillos, y pueden alcanzar las 5000 líneas por minuto.

Las impresoras de tren y de cadena fueron las predecesoras de las impresoras de banda. En realidad, los caracteres en vez de llevarse en una banda continua, eran láminas individuales.

6.2.5 Impresoras de matriz de puntos

Hasta ahora, hemos discutido las impresoras "Solid font", en las que cada carácter se imprime presionando un elemento con la figura del carácter a imprimir, sobre una cinta entintada y el papel. El conjunto de caracteres está limitado por el tipo de elementos instalados en la impresora.

En la figura (6.6) se muestra la cabeza de una impresora de matriz de puntos. Estas impresoras no tienen predeterminado el conjunto de caracteres. En realidad, cada carácter está formado por un patrón de puntos, seleccionados de un array de puntos, los cuales cubren el área asignada al carácter. Si, por ejemplo, la matriz es de 9 puntos de alto y 7 puntos de ancho, la matriz es impresa por una columna de nueve pines (algunas veces llamados agujas). El conjunto de martillos puede adoptar cualquier combinación de pines que sea necesaria para imprimir los puntos necesarios en la primera columna. La cabeza se mueve a la posición de la segunda columna, y el conjunto de martillos adopta la combinación adecuada para imprimir esta segunda columna, y así hasta las siete columnas de la matriz que se hayan impreso. Al igual que en las impresoras de 'Solid font', hay una cinta con tinta entre el papel y los pines. Aquí, la tinta incorpora un lubricante para que los pines corran libremente por sus guías. La generación de los caracteres se realiza de forma similar a la mostrada en el capítulo anterior para dibujar caracteres en una pantalla de vídeo.

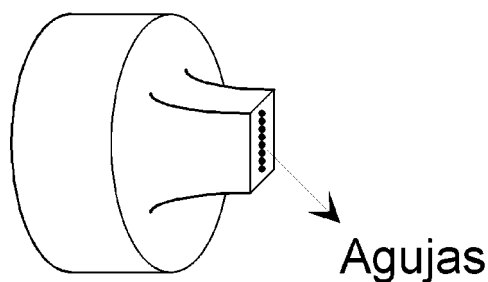


Fig. 6.6 Cabezal de una impresora de matriz de puntos

El número de puntos por posición en la matriz varía de una impresora a otra. Quizás la matriz más común es la de 9 puntos de alto y nueve puntos de ancho para cada carácter. Esto es suficiente para producir caracteres legibles pero no muy elegantes. En muchas impresoras, para prevenir esta restricción, coloca dos columnas sucesivas de pines. Algunas impresoras reducen la altura a 7 u 8 puntos acostando la bajada de las letras 'bajas' tales como 'y', 'p' y 'q'. Unas pocas igualmente modifican estas letras para que no tengan bajada, pero esto hace textos difíciles e incómodos de leer por lo que raramente se da.

Los caracteres impresos son más legibles cuantos más puntos tenga la matriz. Columnas extra pueden ser impresas reduciendo la distancia de los movimientos de la cabeza, después de que cada columna sea impresa. Sin embargo, la velocidad a la que los pines pueden ser conducidos está limitada, por lo que el tiempo de impresión de un carácter se incrementa. Sumarle filas a la matriz no es simple. Un método común es imprimir cada línea de caracteres dos veces, moviendo el papel entre pasos la mitad del espaciamiento de los pines de la impresora. La combinación de los pines en el segundo paso no necesita ser la misma que en el primer paso; esto permite mejorar la figura del carácter. Sin embargo, esto acorta la velocidad de impresión total. Muchas impresoras ofrecen una opción entre estos modos (descritos como 'near letter quality' o NLQ) y el modo borrador simple. Como en las impresoras de margarita, algunos incrementan su velocidad imprimiendo líneas alternativas en direcciones opuestas.

Otro método para incrementar el número de posiciones de puntos sin el requerimiento de la segunda pasada, es incrementar el número de pines en la cabeza de impresión. Los pines pueden ser situados en dos columnas escalonadas por lo que las filas de puntos son impresas de forma entrelazada. A menudo el número de pines es 24 y esto puede producir caracteres de la calidad de las impresoras de margarita, aunque la impresora de matriz no puede conseguir la mayor calidad de impresión porque no puede utilizar cintas de un sólo uso. Las impresoras de 24 agujas usualmente también ofrecen usualmente el modo borrador, con unas pocas columnas por carácter pero una mayor velocidad de impresión. Las impresoras disponibles pueden tener un rango de 7x7 a 36x50 puntos por carácter. Los más populares son los de 9 y 24 agujas, y se han llegado a comercializar hasta con 48.

La selección de los puntos con que se va a imprimir cada carácter, y de este modo la figura del carácter, está determinada por una tabla almacenada en memoria ROM dentro de la impresora. Su conjunto de caracteres es decidido por el diseñador de la impresora, al igual que en las impresoras 'Solid font'. Sin embargo, aquí es fácil cambiarlos con sólo cambiar el chip de ROM (o conjunto de chips) por otro. En algunas impresoras, varios conjuntos de caracteres pueden estar implementados y pueden seleccionarse por el programa conductor de la impresora o por ambos. Otra solución popular es un módulo en la caja de la impresora en el que el usuario puede colocar un cartucho conteniendo la ROM que define el conjunto de caracteres y sus opciones.

Algunas impresoras de matriz también permiten que el carácter sea cargado desde el ordenador host y almacenado en memoria RAM de la impresora. Esto es más útil cuando unos pocos caracteres especiales reemplazan algunos del conjunto estándar, por ejemplo en palabras técnicas o ciertos lenguajes extranjeros. Esta facilidad no se usa muy a menudo puesto que el conjunto estándar de caracteres es bastante extenso e incluye tanto todos los caracteres estándar ASCII como el 'conjunto gráfico extendido de IBM' aunque a veces, éste último sufra pequeñas variaciones.

La mayoría de las impresoras de matriz pueden imprimir también variantes de caracteres de su conjunto estándar. Negrita (producido por múltiples pasadas, como en las impresoras Solid-font) y caracteres subrayados son los más o menos estándar; Itálica es también común. Algunas impresoras, no necesariamente las más caras, ofrecen variantes tales como 'shadow' (sombra) y 'outline' (contorno), y también pueden imprimir caracteres en doble o cuádruple tamaño. El espaciado de los caracteres es a menudo seleccionable, generalmente 10, 12 y alrededor de 16 caracteres por pulgada, y algunas veces también es posible seleccionar el espaciado proporcional y por consiguiente la figura del carácter es ajustable.

Las impresoras de matriz pueden ofrecer una mayor variedad de caracteres que las impresoras 'Solid-font'. Además, este tipo de impresoras pueden trabajar en modo gráfico, en cuyo caso la página no es tratada como un conjunto de caracteres sino como una matriz de pixels, que pueden ser controlados individualmente por el programa, pudiéndose imprimir cualquier imagen.

Un programa puede por supuesto usar el modo gráfico para imprimir caracteres, pero las figuras y tamaños de éste se definen ahora por el programa y por lo tanto no hay restricción de número y forma de ellos. Este rasgo es usado por programas que imprimen pancartas o posters de distintas anchuras y longitudes.

El coste que hay que pagar para usar el modo gráfico es la mayor lentitud de la impresora. Esto no se debe a la impresora en sí misma, ya que no se toma mayor tiempo para imprimir cualquier conjunto de puntos que represente. Sin embargo es necesaria mucha más información para definir una página en modo gráfico que en modo carácter, puesto que un bit describe cada punto mientras que en modo carácter un byte describe el carácter completo. La velocidad con la que la información puede ser pasada a la impresora está limitada por la interfaz entre la impresora y el 'host'. A menudo, sin embargo es más restrictiva la velocidad con la que el programa genera la imagen.

El número de pixels de la página depende tanto del número de agujas de la cabeza como de la distancia en la que la cabeza se mueve horizontalmente entre cada columna de martillos, y la distancia en que la cabeza es movida verticalmente después de que una línea haya sido impresa. Lo último no es necesariamente un múltiplo de la altura cubierta por el conjunto de pines, porque las líneas pueden ser entrelazadas para dar mejor definición. De este modo, no porque la cabeza tenga mayor número de pines, la imagen tendrá mayor resolución, aunque sí puede imprimir más rápido.

La velocidad de modo carácter varía entre alrededor de 40 y 500 caracteres por segundo, dependiendo del precio y de la calidad de impresión, pero 150 caracteres por segundo en modo borrador y 50 en modo NLQ son valores típicos.

Las impresoras de matriz son más versátiles que las impresoras 'Solid-font'. Su fabricación es más barata que la de las impresoras de margarita y además son menos ruidosas y más rápidas que éstas, aunque no son tan rápidas como las de línea.

6.2.6 Impresoras de matriz de líneas

Las impresoras de matriz que hemos descrito anteriormente son impresoras de una línea de un carácter cada vez. Sin embargo, para un mayor rendimiento hay otro tipo de matrices de impacto, las cuales no imprimen carácter a carácter.

En este tipo de impresoras, los pines están dispuestos horizontalmente y espaciados a lo largo de toda la línea en intervalos iguales, típicamente media pulgada. Los pines y los mecanismos que lo conducen están montados en una carcasa o lanzadera "shuttle" (Fig. 6.7), y éste se mueve paralelamente a la línea de impresión, pulsando las agujas cuando estén enfrentadas al lugar donde se debe imprimir. Entonces, el papel se mueve hacia arriba un tanto correspondiente al espaciado vertical de los puntos y el proceso se repite, con la lanzadera moviéndose en la dirección contraria. Este proceso se repite durante toda la página.

Las impresoras de matriz lineal son mucho más rápidas que las impresoras de matriz convencional. Una máquina típica puede imprimir unas 900 líneas por minuto en modo borrador y la mitad de este valor en modo NLQ. Aunque no sean tan rápidas como las impresoras de banda, la posibilidad de variar las fuentes y realizar gráficos hace que las impresoras de matriz lineal sean más útiles para muchas aplicaciones.

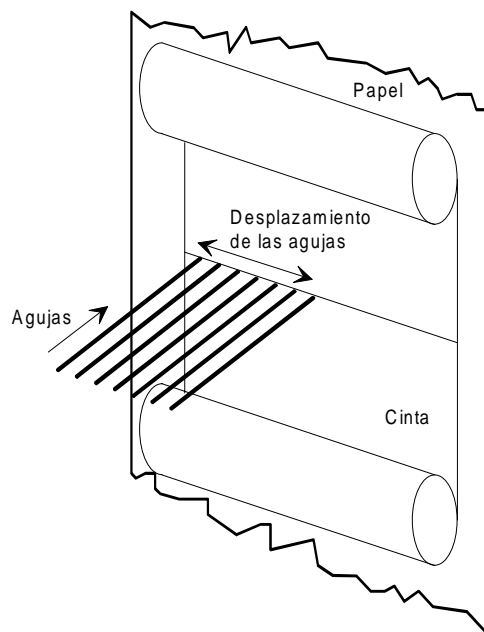


Fig. 6.7 Mecanismo de las impresoras de lanzadera "shuttle" o de matriz lineal

6.2.7 Impresoras de color de matriz

Los caracteres pueden imprimirse escogiendo dos colores usando una cinta con los dos colores, como usan las máquinas de escribir, y algunas impresoras 'solid-font'.

Sin embargo, una vez que la impresora está en modo gráfico, el uso del color se vuelve mucho más atractivo, especialmente si los tres colores básicos pueden ser impresos de forma combinada para dar una mayor gama de colores. Las impresoras de matriz actuales pueden imprimir generalmente en seis colores más el negro y el blanco. Algunas de estas son un poco más caras que las de un sólo color además su mantenimiento es algo más costoso, puesto que la vida de la cinta es más corta.

Entre las impresoras hay detalles que varían, pero el principio básico es que cada línea (si es carácter o gráfico) se imprime tres o cuatro veces: una vez en cada color básico, y algunas veces también una en negro. En cada color requerido sólo se pintan los puntos relativos a ese color. Las cintas tienen tres colores intermedios y cada color se obtiene sobreimpresionando dos o tres colores básicos uno sobre otro. Por ejemplo, el verde sobreimpresionando cyan sobre amarillo. Los colores básicos más comunes en las impresoras son: cyan, magenta y amarillo, los cuales contrastan con la técnica visual donde se usan los verdaderos colores primarios: rojo, azul y verde. En principio, imprimiendo los tres colores primarios debe obtenerse el negro, pero en realidad se consigue el marrón difuso, por lo que las cintas además tienen una banda de color negro por si la impresión no requiere colores.

El color producido por las impresiones de matriz de impacto no es muy bueno, y tiende a hacerse peor a medida que la cinta envejece, debido a que la reutilización de la cinta no hace un uso homogéneo de los distintos colores. Además, la tinta puede ser llevada de una banda de color a otra distinta por los pins. Estos problemas reducen la vida útil de la cinta, que ya es corta puesto que las cintas se pasan tres o cuatro veces por la línea de impresión. Por estas razones, las impresoras de matriz de impacto de color no fueron muy usadas.

6.3 IMPRESORAS DE NO IMPACTO

Algunas impresoras de no-impacto han estado disponibles desde hace bastante tiempo, pero no han sido muy utilizadas debido a su alto costo y problemas tales como el deterioro de la imagen o la imposibilidad de usar papel de 'calco' para producir múltiples copias. Sin embargo, con la llegada de las impresoras láser y otras máquinas de altas prestaciones, las impresoras de no-impacto se han hecho muy populares.

6.3.1 Impresoras de chispa electrostática

Los métodos electrostáticos para marcar en papel fueron usados antes de la existencia de los ordenadores. Necesitaban un papel preparado especialmente, en el que la cara base, generalmente negra tenía una delgada superficie metalizada sobre ella. El papel pasa bajo un rodillo conectado a tierra y bajo una aguja de metal. Cuando se aplica un voltaje a la aguja, se produce una chispa, la cuál pincha la superficie metalizada y por lo tanto puntea la hoja negra.

Los caracteres están grabados. El papel se mueve continuamente y la aguja se mueve a través de él para representar la traza de la señal. Cuando este método se adopta para imprimir caracteres, éstos se forman por un patrón de puntos al igual que ocurría con las impresoras de impacto de matriz. El movimiento de la aguja se reemplaza por una fila de agujas, y éstas se accionan cuando el papel pasa bajo ellas para construir el carácter.

Las impresoras electrostáticas son simples, compactas y virtualmente silenciosas, y la grabación es permanente. Sin embargo, aquí el papel es caro. Estas impresoras se usan sólo para algunas aplicaciones, normalmente donde el papel es estrecho (con lo que sólo se necesitan unas pocas agujas) y su uso es intermitente. Un ejemplo es la impresión en la solicitud de entradas de teatro.

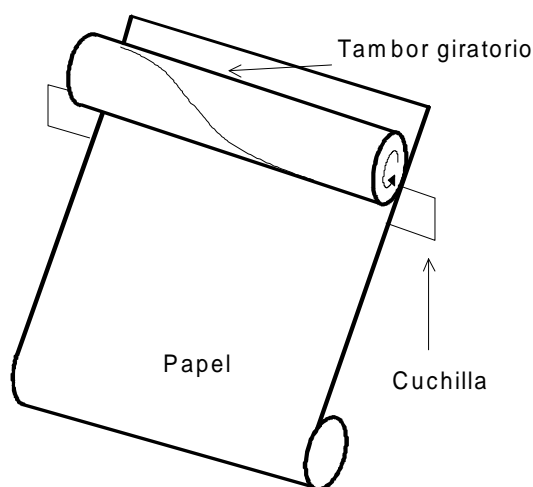


Fig. 6.8 Mecanismo de las impresoras electroquímicas (tipo hélice)

6.3.2 Impresoras electroquímicas

Este método también se introdujo antes que los computadores, en receptores de facsímil ('fax'). Aquí como en el método de la chispa electrostática aparecen puntos visibles donde pasa la corriente a través de un papel preparado. En este caso, el papel es blanco y está impregnado por un producto químico, de tal manera que al pasar una corriente eléctrica a través de él se produce un punto negro. El voltaje necesario para producir la corriente es bastante bajo. Para usarla es necesario mantener el papel húmedo, pero ahora se está usando papel seco. Estas impresoras son prácticamente silenciosas.

En las máquinas de fax, el papel pasa entre un metal perpendicular (Fig. 6.8) como el filo de una navaja y una hélice giratoria, como la paleta de un cortacésped. El punto de contacto (a través del papel) sigue una serie de líneas paralelas a través del papel, como en las pantallas de CRT. El dispositivo puede por lo tanto reproducir imágenes gráficas. Este mecanismo obviamente puede adaptarse para imágenes generadas con un computador, y se ha usado alguna vez para este propósito. Sin embargo, tales impresoras son lentas y las imágenes se despintan con el tiempo. No son muy utilizadas en aplicaciones con el computador.

6.3.3 Impresoras térmicas

Estas son impresoras de matriz muy similar en su diseño a las impresoras de matriz de impacto excepto que la fila de agujas y martillos en la cabeza de la impresora son reemplazados por una fila de elementos de calefacción muy pequeños. La impresora puede ser usada en modo térmico directo con un papel térmico especialmente preparado, que es inicialmente blanco, pero se pone negro cuando se calienta por encima de cierta temperatura. Este tipo de papel se usa ahora bastante en receptores fax. Los elementos calefactores forman un patrón de puntos de la misma forma que las agujas de las impresoras de impacto de matriz. Los elementos han de ser fabricados pequeños para permitir unos 24 elementos por cabeza, por lo que la definición es comparable a la de las impresoras de impacto, y varios estilos de letras (al menos 'borrador' y 'NLQ') pueden estar disponibles. Imprimiendo son virtualmente silenciosas. Como inconvenientes podemos señalar que la imagen no es completamente permanente y el papel térmico cuesta tres o cuatro veces más que el papel normal.

También es posible usar las impresoras térmicas en procesos de transferencia térmica, usando papel blanco y una cinta especial. La cinta de transferencia térmica tiene un revestimiento de cera que se derrite cuando se calienta y es transferido al papel. Esto da una imagen permanente, y usualmente más negra que una de papel térmico. Tiende a tener una superficie brillante como los lápices de cera. Cada cinta se usa una sola vez, por lo que el costo de la cinta sube el precio de grabación del papel (es necesario usar un papel de una superficie muy lisa). Algunas impresoras térmicas son fabricadas exclusivamente para un proceso u otro. Pero muchos pueden usarse con papel térmico sin cinta o papel normal con una cinta de transferencia térmica.

Como con las impresoras de impacto, la impresión de color requiere unos pequeños cambios en la impresora. La cinta en lugar de ser toda negra, lleva tres colores básicos. Esta es generalmente arreglada como una secuencia de sectores, siendo cada sector capaz de imprimir una línea completa. Al final de cada paso, la cinta avanza al comienzo del próximo sector de color. Estas impresoras dan un mejor color que las impresoras de impacto, particularmente porque la cinta se usa una sola vez. No puede obtener un buen color negro imprimiendo los tres colores básicos, por lo que es necesario un sector negro en la cinta. Por otro lado, como la cinta no es reutilizada, su uso es bastante caro.

Las impresoras de matriz térmicas no son muy caras y su rendimiento es similar a las impresoras de matriz de impacto lentas.

Una modificación de las impresoras térmicas de transferencia usa el proceso llamado tinte por sublimación. Aunque son más complejas y caras, nos permiten variar la intensidad de cada color primario y por tanto dan mayor rango de colores, llegando casi a calidad fotográfica pero son muy caras y su coste de mantenimiento es muy elevado.

6.3.4 Impresoras electrográficas

Las impresoras de no impacto de alto rendimiento son generalmente impresoras de páginas, esto es, se introduce una página completa desde el host al buffer de la impresora, y entonces se imprime en una operación. El tipo más importante es la impresora electrográfica o electrofotográfica que a veces es descrita como impresora electrostática (el principio básico es en

realidad electrostático, aunque la tecnología es bastante diferente a la de las impresoras de chispa electrostática descritas anteriormente).

La primera de estas impresoras fue la 'Xeronic' diseñada por Xerox Corporation en 1910 usando la electrografía (o 'xerografía'), principio que acababa de ser introducido por la compañía para las máquinas fotocopadoras; y por supuesto es el más usado en las fotocopadoras de hoy día. El proceso depende de hecho de ciertos materiales, tales como el selenio, el cual almacenará una carga eléctrica mientras permanezca a oscuras, pero se descargará por la incidencia de la luz. De este modo, si la imagen está brillando bajo una película de este material, la carga será retenida en las partes oscuras de la imagen y se perderá en las partes claras. La superficie es entonces explorada con unos polvos aislantes negros (toner), el cual agarra donde la película está cargada, pero falla en cualquier otro sitio. Cuando la hoja es presionada, el polvo de la superficie se transfiere a ella. El papel se calienta para derretir el polvo y fusionarlo al papel, dando una imagen permanente.

La principal diferencia entre la fotocopadora y la impresora es el proceso por el que se forma la imagen. En la fotocopadora esto se hace iluminando el documento original y focalizando su imagen en un tambor mediante un sistema de lentes. En las impresoras Xeronic, una pantalla de tubo de rayos catódicos reemplaza el documento; la información que va a ser impresa se muestra en la pantalla de la misma forma que una pantalla VDU.

El papel se mueve continuamente a través de la impresora, y el rendimiento es alto, mucho más que en los modelos anteriores. La potencia de calentamiento es tan grande que el papel puede quemarse si el calentador permanece encendido mientras el papel está parado. Se usó el papel continuo, y se cortaba por la máquina cuando detectaba las marcas de página en el papel.

La impresora Xeronic era muy compleja y cara, pero su rendimiento era mayor (varias páginas por segundo) que las demás impresoras de su tiempo. En la mayoría de los casos fue utilizada off-line; es decir, los datos que iban a imprimirse eran escritos por el host en un fichero de una cinta magnética, la cinta se llevaba a un conductor de cinta, el cual sólo estaba conectado a la impresora.

La impresora Xeronic ha sido reemplazada por las impresoras láser. Estas trabajan con un principio similar, pero es conveniente describirlas en un apartado separado.

6.3.5 Impresoras Láser

Las impresoras láser (Fig. 6.9) utilizan el mismo método que las impresoras electrográficas, como la Xeronic. La única diferencia en principio es que en lugar de usar un rayo de electrones para formar la imagen en la pantalla de CRT, se usa un rayo de luz formado por un láser que forma la imagen directamente en la superficie del tambor. Usualmente se utiliza un láser semiconductor pero en algunas impresoras de alto rendimiento se usan láseres de gas. La imagen se construye con un rayo controlado para generar puntos brillantes y oscuros. Algunas impresoras láser pueden manejar niveles intermedios de gris. El rayo es deflectado a lo largo de una línea (paralela al eje de abscisa del tambor) inclinando un espejo; y la rotación del tambor proporciona el barrido vertical. El espejo inclinado es de hecho una serie de superficies reflectantes en un tambor giratorio.

La mayoría de las impresoras láser están diseñadas para colocarlas encima de una mesa y usar hojas sueltas. Son impresoras de página, por lo que la memoria propia de la impresora debe almacenar al menos una página completa. El mecanismo puede imprimir con una definición muy alta, a menudo 300, 600 o 1200 puntos por pulgada, y a veces incluso más. Una imagen de una página completa a 400 puntos por pulgada puede necesitar alrededor de 16 megabytes para definirla. La impresora debe tener este buffer de almacenamiento, que es caro, y la interfaz entre el host y la impresora (o a veces software del host) restringe el rango de datos, por lo que la

impresión en modo gráfico es lenta. Esto no es problema en modo carácter donde los símbolos están definidos en ROM (o en RAM cargada por el host) dentro de la impresora.

Como todas las impresoras de impacto, las impresoras láser no pueden hacer copias de carbón. Sin embargo, como la página completa se encuentra en memoria, puede repetirse su impresión todas las veces que se quiera sin tener que ser retransmitida sobre la interfaz. La impresión de copias repetidas es por tanto más rápida que la impresión de la primera copia. Las impresoras láser fueron diseñadas originalmente para usos profesionales, pero la reducción de costes y su alta calidad las ha popularizado siendo el tipo de impresora de calidad más habitual para trabajos que no precisan color y pueden imprimir de 6 a 10 páginas por minuto, aunque existen impresoras de mayores prestaciones capaces de imprimir más de 100 páginas por minuto.

Aunque no muy caras en su rendimiento, las impresoras láser no son baratas de mantener. No necesitan cintas o un papel especial, pero el tóner de fotocopiadora es caro y el tambor necesita ser reemplazado regularmente. Las impresoras recientes usan tambores (ocasionalmente "correas flexibles") revestidos de componentes orgánicos. Estos son más baratos y menos tóxicos que los antiguos con revestimiento de selenio.

Las impresoras láser, como las fotocopiadoras pueden usar tóner de otros colores distintos al negro, aunque normalmente sólo utilizan uno a la vez. Es posible diseñar un mecanismo que solape tres imágenes en tres colores, y esto produce el color a imprimir. Ya están disponibles fotocopiadoras en color. Sin embargo, todavía han aparecido pocas impresoras láser a color, esto puede deberse a que la impresión de colores se dejará para otro tipo de impresoras, tales como las de chorro de tinta o las de sublimación de ceras.

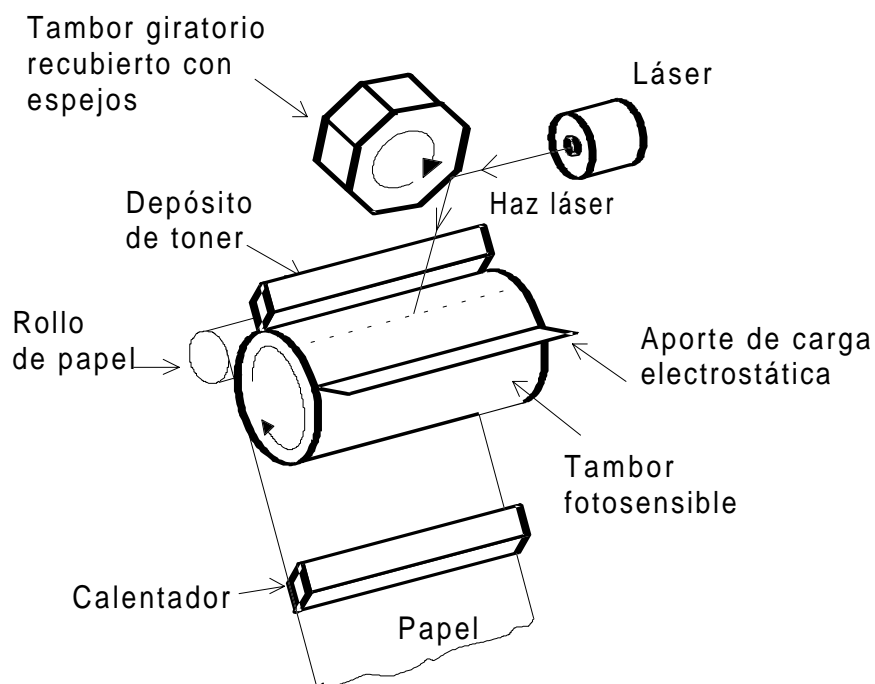


Fig. 6.9 Mecanismo de impresión de las impresoras láser

6.3.6 Impresoras LED, LCD y de deposición de iones

La primera de estas impresoras electrográficas trabaja con un principio similar al que usan las láser; construye la imagen de la página de tal forma que la imagen esté en el tambor fotosensitivo. Sin embargo en lugar de formar cada línea de rastreo con un rayo láser modulado, se forma directamente emitiendo luz de una serie de celdas, una por cada pixel de la línea. En las

impresoras LED cada celda es un semiconductor LED (Light Emitting Diode), similar al usado en algunos display aunque más pequeño. En las impresoras LCD cada celda es un obturador; usando tecnología LCD (cristal líquido) que controla la luz que pasa a través de ella desde una lámpara montada detrás del array LCD. Las características LCD y LED son en general similares a las de las impresoras láser.

Las impresoras de deposición de iones trabajan de la misma forma, pero cada una de las celdas emite un rayo de iones cargados en lugar de un rayo de luz, y el tambor no es sensible a la luz. El tambor tiene una gran superficie de aislamiento, por lo que se puede transferir el toner a un papel sin usar la técnica por calor. De aquí se obtiene que la vida del tambor es infinita y por tanto las impresoras son más baratas de mantener, pero son más caras a la hora de fabricarlas. Por ello son utilizadas en aplicaciones de gran volumen. Pueden imprimir sobre 75 páginas por minuto.

6.3.7 Impresoras magnetográficas

Este tipo de impresoras de nuevo trabajan de una forma similar a las impresoras electrográficas, pero en este caso el tambor está recubierto por una capa magnetizable como los discos magnéticos, y la información se escribe sobre él por una cabeza magnética al igual que en los discos. El toner en polvo seco, que es magnetizable es atraído a la superficie. Las impresoras magnetográficas tienen muchas de las características de las láser, pero la tecnología es difícil y existen pocos dispositivos en el mercado.

6.3.8 Impresoras de inyección de tinta

Son impresoras de matriz y su principio de funcionamiento es simple. En lugar de agujas, como las impresoras de impacto, tienen un conjunto de boquillas y una salida de tinta por cada una para formar un punto en el papel. Los caracteres o las imágenes se construyen por patrones de puntos como en cualquier impresora de matriz. Existen dos técnicas básicas para impulsar la tinta fuera del cabezal de impresión. Por una parte está el sistema empleado en las impresoras de Hewlett Packard y otros fabricantes, en las que un pequeño elemento calefactor produce la ebullición de la tinta generándose una burbuja que empuja una pequeña gota hacia el exterior. Por otra parte tenemos la tecnología liderada por Epson en la que emplean un pequeño elemento piezoeléctrico que empuja a la tinta de forma mecánica. Las prestaciones de ambas tecnologías son similares, pero los compuestos empleados para las tintas deben ser distintos.

La calidad de este tipo de impresoras es excelente, especialmente cuando se utiliza el color. Su principal inconveniente es que la calidad y el realismo depende fuertemente de las características del papel y los papeles que ofrecen buenos resultados resultan excesivamente caros.

Una variante de este esquema es el Continuous-Set o Synchronous Printing. En este caso, cada boquilla tiene una caída continua de tinta en cada posición de pixel; pero las caídas no necesarias se desvían con un campo eléctrico y se conducen a unos canalizadores que la recogen y devuelven al depósito de tinta.

6.3.9 Plotters de plumas

Los plotters pueden agruparse con las impresoras puesto que ambos periféricos pueden representar información en el papel. Sin embargo, mientras que las impresoras construyen la página sistemáticamente por la impresión de puntos individuales o figuras de caracteres predeterminadas, los plotters dibujan la imagen línea a línea, moviendo una pluma a través de un papel como lo haría un delineante. Hay dos tipos básicos de plotters: el 'flat-bed plotter' y el 'drum plotter'. En el primer tipo, el papel se coloca sobre la base del plotter y la pluma se mueve en dos dimensiones, de manera que puede llegar a cualquier punto del papel. En el segundo tipo el papel se fija sobre a un tambor, o simplemente se sitúa sobre él y se fija en una posición por la presión de unos rodillos (como en la máquina de escribir). La pluma se mueve en una única dimensión y el

papel es movido en uno u otro sentido para lograr alcanzar la segunda dimensión. En los dos casos, la pluma puede ser levantada o bajada, por lo que puede moverse sin dibujar nada.

El movimiento en las dos direcciones se controla independientemente y, por lo tanto, puede dibujarse cualquier línea recta o curva. Las líneas son definidas en términos matemáticos: las líneas rectas, por las coordenadas de cada final y las curvas por expresiones matemáticas más complejas. El software asociado con el plotter interpreta estas expresiones y mueve la pluma y el tambor coordinadamente. Algunos plotters usan motores paso a paso para conducir la pluma y el tambor en pequeños incrementos, por lo que líneas no paralelas a los ejes pueden construirse como escalones muy pequeños. Otros son capaces de movimientos lineales y, por lo tanto, pueden dibujar líneas continuas si se les proporciona el software adecuado.

Los caracteres pueden ser dibujados como una serie de líneas. Sin embargo, puede haber problemas al grabar el programa para definirlos. La mayoría de los plotters ofrecen un conjunto de caracteres estándar, en unos cuantos tamaños útiles, que pueden ser especificados por sus códigos de caracteres como se hacía con las impresoras.

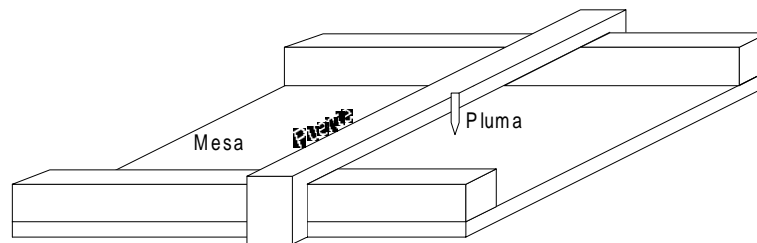


Fig. 6.10 Plotter de plumas plano "flat bed"

Las plumas generalmente son de tipo de punta de bola o de fibra. Están disponibles en un rango de colores, ancho de línea y distintos tipos de tintas para imprimir en papel, transparencias u, ocasionalmente, cristal o metal. Para dibujar con varios colores o anchos de línea pueden usarse diferentes plumas por turnos. En los plotters de bajo coste, la actividad se detiene para que las plumas de color sean cambiadas de forma manual. En plotters de mayor calidad las plumas se cambian automáticamente.

Los plotters funcionan mejor cuando se dibujan líneas. Sin embargo, las áreas de color se rellenan dibujando líneas paralelas, lo cual tiende a ser lento, incluso con plumas que dibujen puntos más anchos. De hecho, los plotters son generalmente lentos aunque, por supuesto, el tiempo que se toman dependerá del número y longitud de las líneas que se están dibujando.

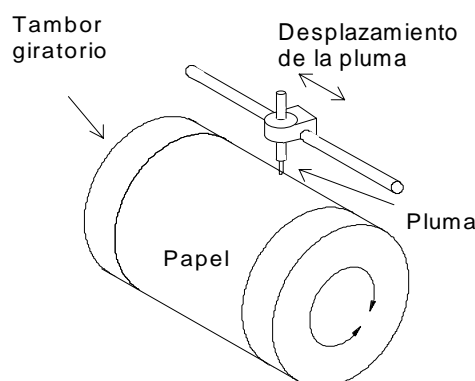


Fig. 6.11 Mecanismo de plotter de tipo tambor, o de papel continuo

Los plotters 'flat bed' (Fig. 6.10) tienen una mejor precisión que los de tipo 'drum' (Fig. 6.11) y son necesarios si se está usando un medio rígido. Sin embargo, se vuelven más caros a medida que el tamaño del papel se va haciendo mayor, por lo que para tamaños por encima de A2 ó A3 suelen usarse los plotters de tambor de forma casi exclusiva, excepto si hace falta una gran precisión o hay que dibujar sobre un medio rígido.

Claramente, el método por el que las instrucciones de software mueven la pluma deberían ser estándar. No existe un estándar universal, pero las especificaciones tienden a ser las mismas que las de las firmas que llevan la delantera en el desarrollo de plotters. La más popular es la de Hewlett Packard, o HPGL ('Hewlett Packard Graphic Language').

Los plotters son tan comunes como las impresoras, teniendo dos importantes campos de aplicación: el primero es el CAD (Computer Aided Design), donde habitualmente ingenieros y arquitectos usan plotters de gran precisión; el otro es en aplicaciones de presentación gráfica, en el campo comercial. Aquí los plotters son usados para dar una imagen mejor de la presentación de caracteres coloreados y gráficas, a menudo como acompañante de un proyector de diapositivas. Para estos propósitos se usan habitualmente los plotters de A4 ó A3, siendo el cambio automático de las plumas esencial. Estos plotters son algo más caros que las buenas impresoras matriciales. Los plotters de precisión son mucho más caros.

En la actualidad, los plotters de plumillas han dejado paso a los de inyección de tinta que emplean la misma técnica que las impresoras. Esto ha hecho que los plotters planos o de reducidas dimensiones hayan desaparecido del mercado en beneficio de este tipo de impresoras. El nombre de ploter ha quedado relegado a los de tambor y de grandes dimensiones y que pueden crear documentos de dimensiones considerables ($>1\text{m}^2$) y trabajar con papel continuo. La diferencia fundamental entre estos plotters y las impresoras de inyección es que emplean un cabezal de impresión permanente y que no es reemplazado a la vez que se cambia el cartucho de tinta. Esto hace que sean cabezales más costosos pero permiten una mayor economía al permitir añadir tinta de cualquier color básico (negro, amarillo, cyan o magenta) sin necesidad de cambiar el cabezal de impresión.

6.3.10 Plotters electrostáticos

Aunque se describen como plotters, son más bien impresoras de matriz; la imagen se construye con un conjunto de plumas y es transferida al papel de la misma forma que en las impresoras láser. Sin embargo, estos plotters tienen controles que aceptan datos en forma de definición de líneas (vectores), rango de caracteres o pixels. Se diseñan para utilizarse con grandes hojas de papel y usan un tambor o mecanismo híbrido. La definición es, típicamente, de 400 puntos por pulgada. Estos plotters pueden ser usados para usar tres colores básicos, como en las impresoras de matriz. Sin embargo, esto requiere el mecanismo triplicado o bien realizar múltiples pasadas, lo cual tiene sentido cuando queremos mucha precisión.

En todo caso, el plotter en sí mismo o el host tienen que convertir la información de vector en forma 'raster'. Esto usa una gran cantidad de potencia del ordenador y suele realizarse por un hardware dedicado mejor que por el software.

Los plotters electrostáticos no pueden alcanzar la resolución de los mejores plotters de pluma, pero como contrapartida son más rápidos. También son más efectivos donde las letras son la parte fundamental, ya que cada letra es formada como una alta definición de patrones de puntos más que como una secuencia de líneas dibujadas.

6.4 DITHERING O ENTRAMADO

Esta es una técnica usada en varios tipos de impresoras, en las monocromas para producir escalas de grises, y en las de color para incrementar el rango de colores. El principio de esta técnica se muestra en la figura (6.12) y consiste en que cada pixel (en el sentido de la menor unidad cuyo color o intensidad es definida por el computador) se representa por un grupo de puntos (normalmente 4 ó 9) formando un array rectangular. Supongamos que tenemos una matriz de 4 puntos: si el conjunto de 4 que tenemos de 0 a 4 puntos pueden ser negros, conseguimos el efecto de tener 4 niveles de gris además del blanco. Si la impresora es de color, cada uno de los tres colores básicos puede tomar 4 intensidades diferentes, con lo cual ampliamos el número de colores.

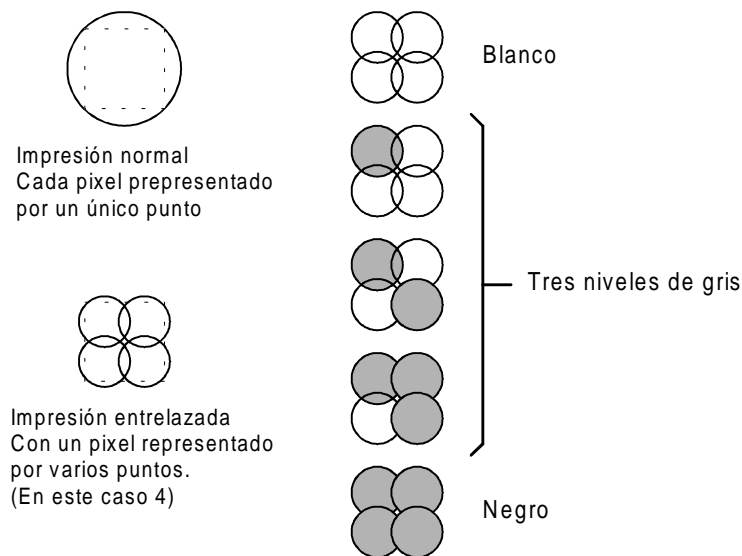


Fig. 6.12 Dithering o entramado

El inconveniente que tiene es que debemos partir por la mitad el espaciado de los puntos en cada dirección (imprimiendo cuatro veces los puntos) o partir por la mitad la resolución. La segunda solución es perfectamente aceptable, por ejemplo, cuando todo el rango de colores no se utiliza para las líneas pero sí se rellenan áreas de líneas dibujadas o caracteres.

Sistemas de instrumentación y control

7.1 TRANSDUCTORES Y SEÑALES DE CAMPO

Transductor es todo dispositivo o elemento que convierte una señal de entrada en una de salida pero de diferente naturaleza física. Normalmente se desea transformar señales de las variables físicas o químicas que deseamos medir, en magnitudes eléctricas que son las que manejamos con más facilidad en instrumentación. La salida del transductor es una función conocida de la magnitud de entrada y la relación entre ambas (magnitud a medir y salida del transductor) puede no ser lineal, aunque se procura que lo sea para simplificar su tratamiento.

Aunque lo más habitual es que una de las dos formas de energía que intervienen en el proceso de transducción sea eléctrica, no siempre es cierto. Pensemos por ejemplo en los micrófonos ópticos en los que el sonido produce deformaciones en una lámina metálica en la que se refleja un haz luminoso. La señal de salida es una variación en el brillo del haz reflejado que posteriormente será convertida mediante un fotodetector (que no es más que un transductor electroóptico) a una señal eléctrica.

La señal eléctrica tal como la proporciona un transductor no es, en general, directamente utilizable por un sistema de adquisición de datos conectado a un ordenador. Por eso suele someterse a estas señales a una serie de procesos típicos. Estos pueden ser entre otros: aislamiento, acoplo de impedancias, cambios de nivel o tipo de la señal, amplificación, filtrado, linealización, cálculos varios (p. ej. $\sqrt{\quad}$), etc. Estos procesos pueden efectuarse en el propio transductor, en el sistema de adquisición de datos o en un punto intermedio.

Uno de los procesos deseados suele ser la amplificación o conversión de la señal al rango de tensión usual en los sistemas de adquisición de datos (0 a 10V); esto puede requerir una atenuación para señales más elevadas, o una amplificación apropiada para las de niveles bajos. Otro es su transformación al rango habitual de corriente en proceso de datos de campo (4 a 20 mA), para poder transmitirlos por cable trenzado a distancia. La transmisión en corriente proporciona una notoria inmunidad al ruido ya que la información no es afectada por caídas de tensión en la línea, impulsos parásitos, resistencias o voltajes inducidos por contaminación electromagnética, etc.

Desde el punto de vista de las señales que proporcionan estos transductores se pueden clasificar en:

- 1) Transductores de resistencia variable
- 2) Transductores de reactancia variable (capacitivos o inductivos)
- 3) Transductores generadores de carga
- 4) Transductores generadores de tensión
- 5) Transductores generadores de corriente
- 6) Transductores digitales

En esta pequeña lista no están incluidos todos los tipos posibles pero sí los más habituales. Los dos primeros son de tipo pasivo (no generan señal, sólo la transforman, el resto se consideran activos (sí generan señal). El hecho de que generen una señal no implica necesariamente que deban ser alimentados de forma externa. Como ejemplo tenemos los transductores piezoeléctricos que generan una tensión entre sus dos extremos, cuando son sometidos a presión o deformación.

Para su introducción en un sistema de instrumentación con osciloscopios digitales o conexión a ordenador, los que generan señal no presentan problemas ya que pueden ser conectados directamente al ordenador. Hay materiales que permiten variar su resistencia como respuesta a casi cualquier fenómeno físico: temperatura, presión, humedad, etc., por lo que la variedad de este tipo de transductores es inmensa.

7.1.1 Transductores de resistencia variable

Son muy populares y se utilizan en la medida de numerosas variables, ya que es la salida de aquellos que utilizan potenciómetros lineales de cursor deslizante, galgas extensiométricas, termómetros resistivos (termorresistencias RTD y termistores), magnetorresistencias, resistencias dependientes de la luz (LDR), higrómetros resistivos, etc.

Para obtener una señal de salida se deben tener en cuenta dos fenómenos, el primero es la necesidad de una alimentación eléctrica ya que la resistencia en sí no genera ninguna señal y el segundo es que esta alimentación influye en la salida por el posible autocalentamiento del transductor.

La medida de la resistencia se puede hacer de forma directa, es decir, como una aplicación de la ley de Ohm midiendo la corriente que la atraviesa a una cierta tensión o la tensión que cae en ella a una corriente constante. Pero el método más usado por ser el más preciso y sensible es el que utiliza un puente de Wheatstone. Sobre este tipo de medidas existe una gran bibliografía que se puede encontrar en cualquier texto de instrumentación. Su salida se realiza a través de un amplificador diferencial que proporciona una señal en tensión, que es la más usada como entrada de un sistema de adquisición de datos conectado a un ordenador personal.

7.1.2 Transductores de reactancia variable (capacitivos o inductivos)

Los transductores capacitivos son muy usados cuando se quiere detectar desplazamientos muy pequeños (hasta 10^{-9} cm.), ya que poseen una gran estabilidad y precisión. También se utilizan para medida de niveles de líquidos conductores o dieléctricos, medida de espesores de dieléctricos, etc. Los transductores inductivos son muy usados ya que se incorporan en muchos equipos que los usan como transformadores de desplazamientos en señales eléctricas. Se suelen dividir en tres grupos principales: los de reluctancia variable, los de corrientes de Foucault y los transformadores diferenciales (LVDT).

La medida en estos transductores se debe realizar en alterna y por lo tanto a continuación, deberá haber un sistema de conversión de alterna a continua, que puede ser de valor eficaz, de valor medio o de pico. La medida propiamente dicha se puede hacer por medio de un divisor de tensión aplicando directamente la ley de Ohm, utilizando un puente de alterna o un oscilador de frecuencia variable. En cualquier caso su paso a tensión continua es necesario para su utilización en un sistema de adquisición de datos por ordenador.

7.1.3 Transductores generadores de carga

En realidad los transductores generadores de carga son generadores de corriente pero en estado de reposo poseen resistencias muy altas y por lo tanto corrientes muy bajas. Son muy usados para medida de radiación, células fotoeléctricas, células de ionización, transductores piezoeléctricos. Su medida depende del transductor y del uso que se desee de la medida. Si se desea una medida continua se utilizan amplificadores, convertidores tensión-corriente o amplificadores de carga. Pero si se desea analizar los impulsos (número, tensión máxima, etc.) deberán utilizarse amplificadores y analizadores de impulsos.

7.1.4 Transductores generadores de tensión

Estos transductores están bastante extendidos. Destacan los termopares, pHmetros, medidores Redox, etc. Además, numerosos equipos que no generan esta salida directamente del sensor, la presentan en su salida por medio de conversiones electrónicas internas. La ventaja que presentan es que no necesitan ninguna acción para su introducción en sistemas de adquisición de datos por ordenador salvo quizás, una adaptación de niveles de tensión. Su desventaja es la transmisión a distancia ya que ésta puede ser afectada por ruidos.

7.1.5 Transductores generadores de corriente

Existen numerosos transductores que presentan salida en corriente, ya que es la salida más extendida en equipos de instrumentación para la transmisión de señales de campo (4-20 mA), por lo que la transformación en tensión de estas señales es una práctica muy generalizada, antes de introducirlas en el sistema de adquisición de datos que suele trabajar en tensión. La conversión corriente-tensión se realiza simplemente usando una resistencia de precisión.

7.1.6 Transductores digitales

Estos transductores son muy utilizados en equipos electromecánicos para indicar acciones, por ejemplo finales de carrera, interruptores de diferentes magnitudes, alarmas, etc. Desde el punto de vista de su introducción al ordenador no presentan más problema que la adaptación de sus niveles de tensión.

7.2 SISTEMAS DE ADQUISICIÓN DE DATOS

7.2.1 Introducción

Los sistemas digitales de control se utilizan ampliamente debido a su bajo coste en comparación con los analógicos. Presentan ventajas en cuanto inmunidad al ruido, precisión y facilidad de implementar funciones complejas. El principal inconveniente es que tienen una respuesta más lenta, aunque para la mayoría de las aplicaciones esto no es un inconveniente. Los

sistemas de control de procesos con realimentación computerizada se utilizan en muchas industrias para controlar sus distintos procesos de fabricación. En el mundo físico, las variables son continuas y es preciso transformarlas, amplificarlas y convertirlas a variables digitales para que un sistema digital las pueda procesar. Los sistemas de adquisición de datos realizan todas estas funciones. En otras palabras, los sistemas de adquisición y conversión de datos se usan para procesar señales analógicas y convertirlas en digitales para su posterior procesamiento o análisis mediante computador o en nuestro caso en un ordenador personal.

En general, un sistema de adquisición de datos toma una magnitud física tal como presión, temperatura, posición, etc. y la convierte en una tensión o corriente eléctrica que será posteriormente muestreada y cuantificada (digitalizada). Una vez conseguido esto, todo el posterior tratamiento de la señal se realiza por circuitos electrónicos digitales.

En principio tiene lugar un tratamiento electrónico y al terminar éste, la señal se convierte en digital mediante un convertidor o conversor A/D (analógico/digital). Esta salida digital puede ir a diferentes sistemas digitales tales como un ordenador, un controlador digital, un transmisor de datos digital, etc.

Un circuito completo de adquisición de datos se indica en la figura (7.1) con todos los componentes fundamentales y sus interconexiones.

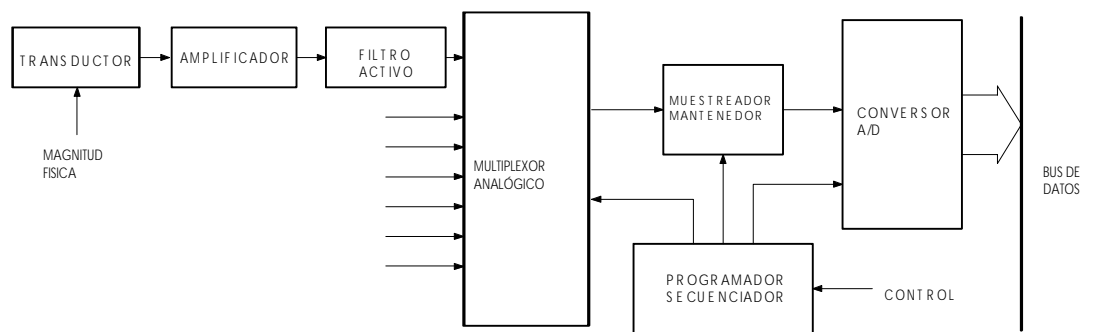


Fig. 7.1 Esquema general de un sistema de adquisición de datos

La entrada al sistema (el parámetro físico a medir), se convierte en una magnitud eléctrica por el transductor y ésta se lleva a la entrada del amplificador. La misión de éste es preparar la señal de salida del transductor al nivel de tensión necesario (1 a 10V) para atacar al siguiente circuito analógico. Sigue al amplificador un filtro activo paso baja, usado para eliminar los componentes de alta frecuencia o ruido de la señal. En ocasiones se puede necesitar hacer con la señal alguna operación no lineal en cuyo caso ésta se puede hacer antes o después del filtrado.

A continuación, la señal va a un multiplexor analógico en el que cada canal de entrada es conectado secuencialmente a la salida durante un periodo de tiempo especificado. De esta forma los circuitos que siguen al multiplexor son compartidos secuencialmente por un cierto número de señales analógicas.

La salida del multiplexor analógico va a un circuito de muestreo y retención ('sample and hold'), el cual muestrea la salida del multiplexor en un momento determinado y mantiene el nivel de tensión en su salida hasta que el conversor (A/D) realiza la conversión.

Por último, la programación y secuencia de tiempos de la operación se realiza por los circuitos de control que a partir de las salidas digitales de control, procedentes del ordenador personal, controla al multiplexor, 'sample and hold' y conversor A/D.

Veamos a continuación algunos principios en los que se basa la conversión analógico-digital de la información.

7.2.2 Cuantificación y codificación

La conversión A/D es en su forma conceptual básica un proceso de dos pasos: cuantificación y codificación.

Cuantificar es el proceso de convertir una entrada analógica continua en una serie de niveles discretos de salida. Estos niveles se pueden identificar por una serie de números, en general como un código binario. La operación de cuantificar una señal se ilustra por la figura (7.2) que muestra la transferencia de las tensiones continuas a valores discretos con ocho estados de salida correspondientes a un conversor A/D de tres dígitos. Los ocho estados binarios tienen asignada la secuencia de números binarios desde el 000 al 111. El número de estados de salida para una codificación binaria de un convertidor A/D es 2^n donde n es el número de bits. Por lo tanto, un convertidor de ocho bits tendrá 256 estados de salida y uno de 12 bits, 4096.

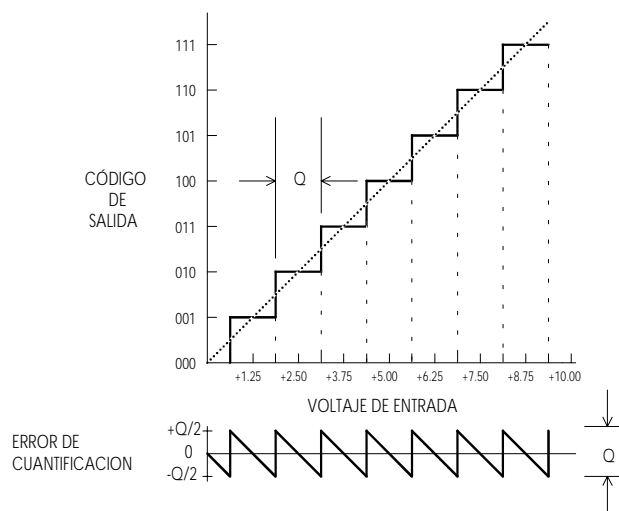


Fig. 7.2 Cuantificación de una señal continua

Esta función cuantificadora tiene algunas características importantes:

- Su resolución, que es el número de estados de salida expresados en bits (en este caso, 3 bits).
- Los niveles de decisión analógica o niveles de umbral; en el caso de la figura (7.2), los valores de 0.625, 1.875, 3.125, 4.375, 5.625 y 8.125. Hay $2^n - 1$ puntos de decisión analógica.
- Los niveles de decisión están colocados a medio camino entre el centro de los puntos de las palabras del código y que en el caso de la figura (7.2) corresponden a los valores de tensión 1.25, 2.50, 3.75, 5.00, 6.25, 7.50 y 8.75 V.

La distancia entre los niveles de decisión codificados se expresa por Q (intervalo de cuantificación). Si para todo el rango de variación de la señal analógica de entrada, restamos ésta de la salida (niveles discretos), obtendremos una señal de error. Este error llamado error de cuantificación es intrínseco del proceso (no se puede eliminar por tanto) y depende del número de niveles de cuantificación o resolución del cuantificador. La salida por tanto se puede considerar como la entrada analógica con un ruido (el de cuantificación) asociado a ella.

Un conversor A/D hace las operaciones de cuantificar y codificar una señal en un tiempo determinado. El tiempo requerido para hacer una medida o conversión se denomina generalmente 'tiempo de apertura' (t_a). La velocidad de conversión requerida en un caso particular depende de la variación temporal de la señal a convertir y del grado de resolución requerido. El tiempo de apertura se puede considerar como una incertidumbre de tiempo (error) en hacer una medida y resulta en una incertidumbre en amplitud si la señal está cambiando durante ese tiempo. Como se ve en la figura 7.3, la señal de entrada al convertidor A/D cambia ΔV durante el tiempo de apertura t_a en que la conversión se efectúa. El error puede ser considerado como un error en amplitud o un error en tiempo. Los dos están relacionados como sigue:

$$\Delta V = t_a \frac{dV(t)}{dt}$$

donde $\frac{dV(t)}{dt}$ es la velocidad de cambio en el tiempo de la señal de entrada.

Si a partir de aquí obtuvieramos el tiempo necesario para digitalizar una determinada frecuencia de señal con un cierto grado de resolución veríamos que para convertir una señal de variaciones relativamente lentas (p. ej. 1 KHz) con una moderada resolución (10 bits), se requiere un conversor A/D extremadamente rápido (tiempo de apertura no superior a 160 nseg.) y por tanto muy caro. Pero este problema se puede resolver de una manera muy simple y barata usando un circuito 'sample and hold', el cual reduce el tiempo de apertura considerablemente al tomar un muestreo rápido de la señal y mantener su valor durante el tiempo requerido para la conversión.

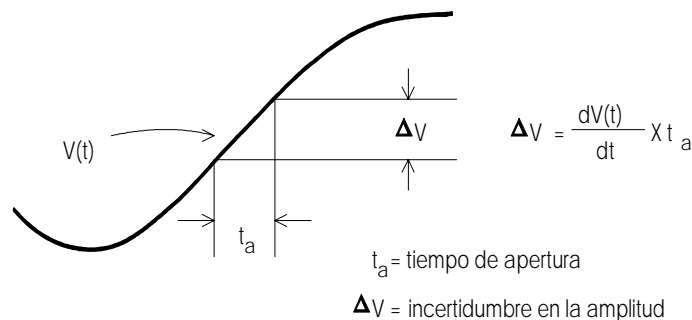


Fig. 7.3 Relación entre el tiempo de apertura y la incertidumbre de amplitud

7.2.3 Muestreo y "aliasing"

La operación de muestreo está indicada en la figura 7.4 en la que vemos una señal analógica (a) y un tren de impulsos de muestreo (b). El resultado del proceso de muestreo es el mismo que obtendríamos al multiplicar la señal analógica de entrada por un tren de impulsos de amplitud unidad. La señal modulada resultante se ve en la parte (c) donde la amplitud de la señal analógica está contenida en la envolvente de los impulsos.

El propósito del muestreo es utilizar de una forma eficiente los equipos procesadores de datos y facilitar la transmisión de los mismos. Un simple SAD (sistema de adquisición de datos), por ejemplo, puede utilizarse para transmitir varios canales analógicos basándose en el muestreo de forma secuencial, con la ventaja respecto al sistema antieconómico de utilizar varios canales de transmisión para enviar continuamente varias señales.

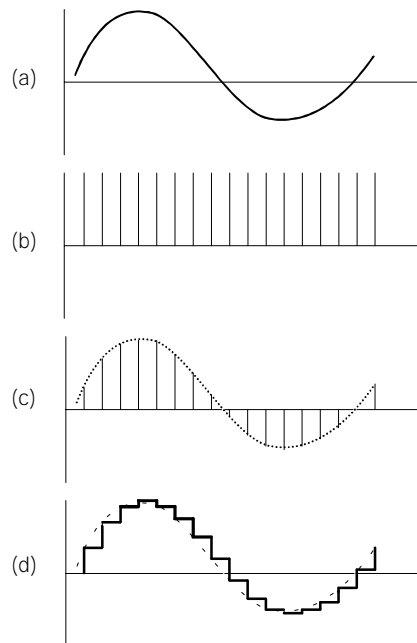
Si la señal analógica es muestreada y memorizada (mantenida) entre los impulsos de muestreo, el resultado es el indicado en la figura 7.4 (d). Este es el trabajo que realiza un circuito llamado de muestreo y retención ('sample and hold'). En los equipos de proceso de datos para vigilancia y control de procesos, puede ser suficiente muestrear el estado del proceso solamente

una vez cada cierto tiempo, realizando el cálculo y corrección oportunos y a continuación liberar el computador para otras tareas.

No se debe olvidar que el objeto de sistemas de conversión de datos es la reconstrucción fiel de la señal a partir de los datos adquiridos. Será necesario saber cada cuanto tiempo se debe tomar una muestra de una señal para no tener pérdidas de su información. Si una señal es lenta, se puede extraer toda su información fácilmente al muestrear de forma que no haya cambio, o éste sea muy pequeño, entre cada muestra. Habrá una pérdida de información si hay un cambio significativo en la amplitud de la señal entre cada muestra. La frecuencia con que se debe muestrear una señal para no perder información de la misma viene dada por el teorema de muestreo ('Sampling Theorem'): "Si el espectro de frecuencias de una señal analógica no contiene componentes de frecuencia superiores a f_c , la señal original puede ser completamente recuperada sin distorsión, si es muestreada a un ritmo de al menos $2f_c$ muestras por segundo".

Fig. 7.4 Operación de muestreo:

- a) Señal analógica a muestrear
- b) Tren de impulsos de muestreo
- c) Señal modulada
- d) Señal muestreada y mantenida



El teorema de muestreo se puede ilustrar con el espectro de frecuencias de la figura 7.5. La figura 7.5(a) muestra el espectro de una señal continua con componentes de frecuencia limitadas por la frecuencia f_c .

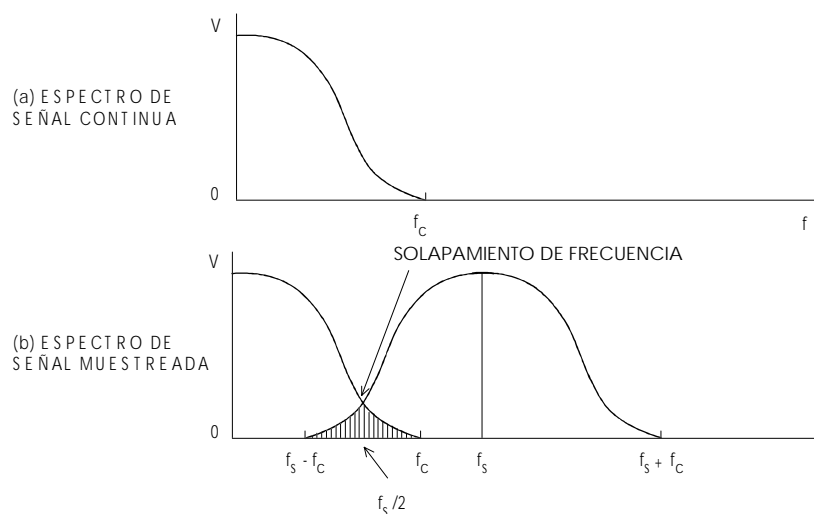


Fig. 7.5 Espectro en frecuencia de la señal muestreada

Cuando esta señal es muestreada a un ritmo f_s el proceso de modulación da como resultado el espectro mostrado en la figura 7.5(b). Aquí debido a que el ritmo de muestreo no es suficiente, algunas de las componentes de alta frecuencia de la señal se pliegan en el espectro. Este efecto es el llamado plegado de frecuencias ('frequency folding'). En el proceso de recuperación de la señal original, las componentes de frecuencias plegadas causan distorsión y no se pueden separar o distinguir de la señal original.

Se elimina el plegado de frecuencias usando una frecuencia de muestreo suficientemente alta o filtrando la señal original para eliminar las componentes de frecuencia mayor de $f_c/2$.

En la práctica no obstante, hay siempre algún plegado de frecuencias debido al ruido y filtros no ideales. Debe tratarse de reducir este efecto a proporciones despreciables.

Otro efecto consecuencia del plegado es conocido como 'aliasing'. La figura 7.6 ilustra esto mostrando una señal periódica que se muestrea a un ritmo menor que dos veces por ciclo. Las amplitudes de muestreo indican unidas por una línea de puntos que evidentemente tiene un periodo bastante diferente de la señal original y es una 'alias'. En esta figura puede verse que si la forma de onda es muestreada al menos dos veces por periodo como requiere el teorema de muestreo, la frecuencia original se mantiene.

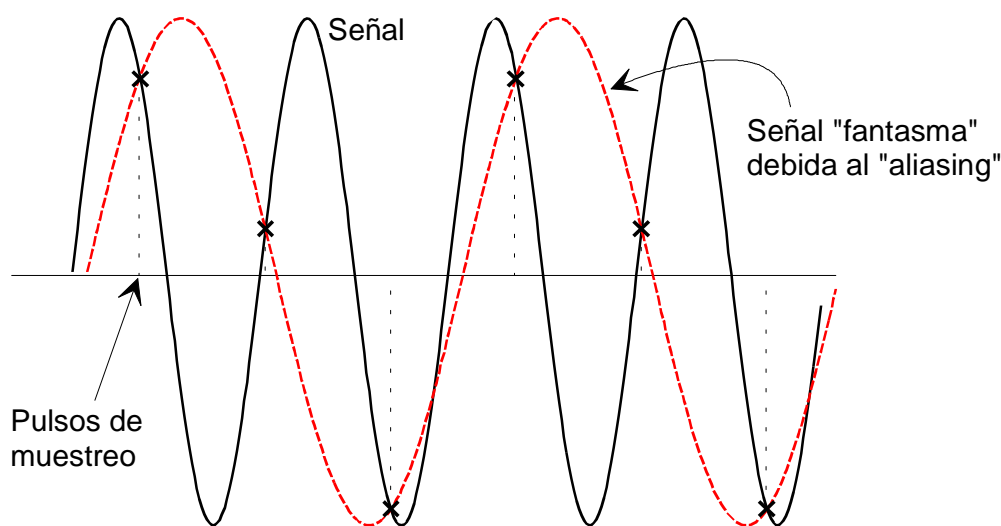


Fig. 7.6 Efecto de aliasing

7.3 CIRCUITOS BÁSICOS DE UN SISTEMA DE ADQUISICIÓN DE DATOS

Describimos a continuación el funcionamiento de los circuitos que componen un sistema de adquisición de datos.

7.3.1 Amplificadores

La primera parte de un sistema de adquisición y conversión de datos trata de extraer la señal a medir. El procesamiento inicial de la señal se hace con un amplificador, filtro y posiblemente un 'operador' no lineal. El propósito del amplificador es realizar una o más de las siguientes tareas: aumentar la amplitud de la señal, adaptar impedancias, convertir una señal de corriente a tensión o separar una señal diferencial del ruido en modo común. En la mayoría de los sistemas de conversión de datos el nivel deseado de tensión de salida es de 5 a 10V a fondo de escala. Este es

el nivel aceptado por la mayoría de los multiplexores analógicos, 'sample and holds', y conversores A/D.

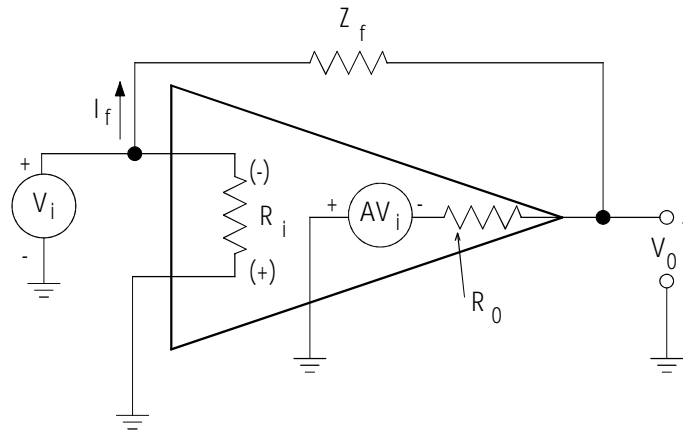


Fig. 7.7 Amplificador operacional ideal

- Impedancia de entrada: $R_i = \infty$
- Ganancia en lazo abierto: $A = \infty$
(sin resistencia de realimentación entre entrada y salida)
- Impedancia de salida $R_o = 0$
- Ancho de banda $BW = \infty$
(rango de frecuencias en el que mantiene las mismas propiedades)

Un elemento imprescindible en todo sistema de adquisición de datos es el amplificador operacional. En la figura 7.7 se muestran las características de un amplificador operacional ideal (A.O.I.). Como puede comprobarse, estas características son imposibles de alcanzar en la práctica pero sirven para un estudio cualitativo de su comportamiento. De hecho un amplificador operacional comercial será tanto mejor cuanto más se acerque a estas características ideales. El diseño y la estructura interna de un elemento de este tipo queda fuera de los objetivos de este curso, aunque conviene saber de su existencia y función. Se puede emplear de forma aislada para adaptar o amplificar señales, o también configurado como comparador dentro de un circuito conversor A/D o D/A. En la figura 7.8 se muestran algunas de las configuraciones más elementales realizadas con A.O.

Después del amplificador puede ser necesario usar un filtro paso baja para reducir la interferencia del ruido sobre la señal y para limitar la anchura de banda de la señal analógica a menos de la mitad de la frecuencia de muestreo. En este último caso se denominan filtros 'antialiasing'. Últimamente un tipo de filtro utilizado cada vez más debido a su facilidad de implementación es el de conmutación de capacidad mediante el cual, con un sólo circuito integrado y pocos componentes externos más, se puede obtener filtros que de otra forma requerirían muchos componentes discretos.

7.3.2 Codificación digital

Los conversores A/D y D/A relacionan los valores analógicos y digitales mediante un código digital apropiado. Los códigos usados son binarios y entre éstos el más común es el binario puro. Un número binario puro se representa como:

$$N = a_n 2^n + a_{n-1} 2^{n-1} + \dots + a_1 2^1 + a_0 2^0$$

donde los coeficientes a_n toman los valores '0' ó '1'.

En un conversor A/D o D/A el primer bit es llamado bit más significativo o MSB ('most significant bit') y tiene un peso de $1/2$ del fondo de escala (FS) del conversor, el segundo bit tiene un peso de $1/4$ de FS y así sucesivamente hasta el último bit llamado bit menos significativo o LSB ('least significant bit').

La resolución del conversor está determinada por el número de bits y el valor de los intervalos o amplitud del LSB viene dado por $FS/2^n$, esto ha sido llamado anteriormente Q (intervalo de cualificación)

El valor analógico del fondo de escala para un conversor puede ser cualquier voltaje conveniente pero intervalos de 0 a 5 V y 0 a +10 V en los de entrada analógica unipolar y de ± 5 V y ± 10 V en los bipolares (ó diferenciales), son los usados más comúnmente.

Para valores analógicos bipolares los códigos más comunes son el binario desplazado ('offset binary') y el complementado a 2 (2's complement). Otro código muy usado en ambos tipos de conversores es el decimal codificado en binario, BCD ('binary coded decimal') en el que 4 dígitos binarios se usan para cada dígito decimal. Esto es muy usado en multímetros y otros aparatos que poseen indicadores de salida.

En ciertas ocasiones se utiliza el código Gray que cambia un solo bit para pasar de un número a su inmediato. Este código reduce la ambigüedad en los casos de medidas consecutivas, es decir, sistemas en los que para ir de un valor a otro, se debe pasar por todos los intermedios; por ejemplo, la determinación de posiciones angulares mediante un disco giratorio.

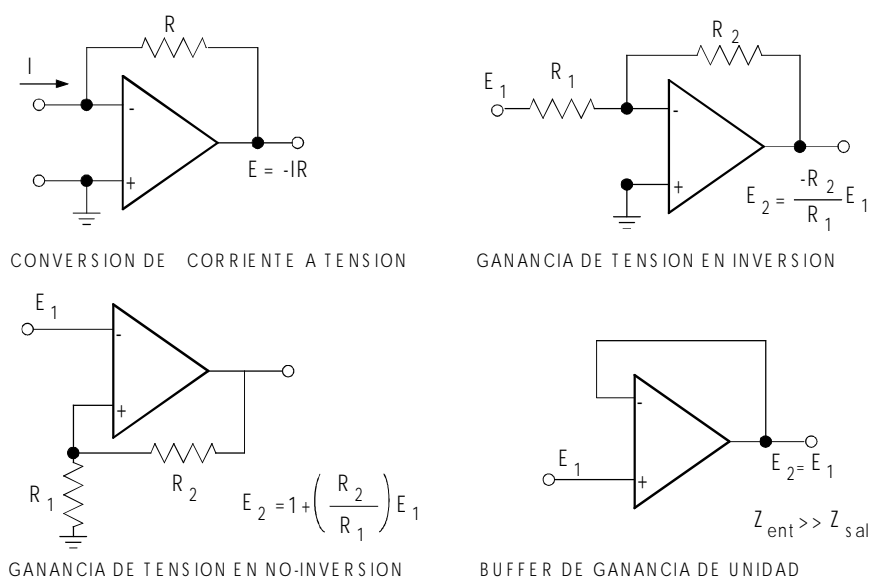


Fig. 7.8 Configuraciones con amplificadores operacionales

7.3.3 Conversores digitales/analógicos (D/A)

Estos conversores son usados en la comunicación del ordenador con el mundo exterior para una gran cantidad de aplicaciones específicas. Además, estos conversores D/A son componentes de gran cantidad de conversores A/D. A continuación veremos su principio de funcionamiento pero limitándonos al método de conversión paralelo que es el más comúnmente usado y cuya configuración básica veremos en la figura 7.9.

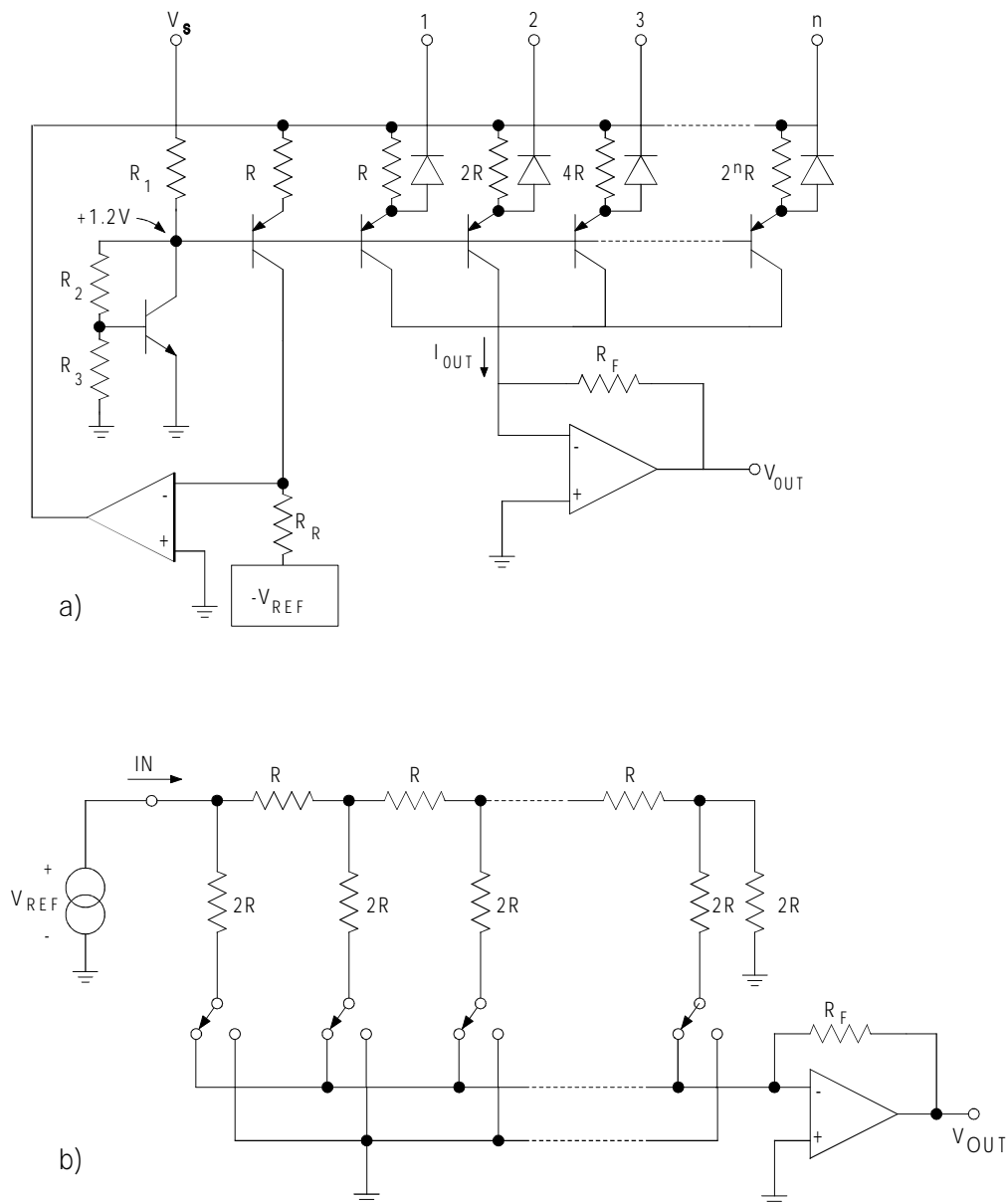


Fig. 7.9 Conversores Digitales-Analógicos

Un interfaz digital convierte las entradas lógicas a los niveles de control de una serie de interruptores. Estos operan junto con una red escalonada de resistencias de precisión, como un sumador ponderado para dar voltajes cuantificados según los pesos binarios 1, 2, 4, etc. La red de resistencias está referenciada a una fuente de tensión precisa y estable (V_{ref}). La salida de esta red es la suma de los 'pesos' binarios en forma de tensión. Existen conversores con salida en corriente que son conversores con salida en tensión con un amplificador operacional que hace de conversor corriente-tensión.

El circuito que utiliza resistencias ponderadas cuantificadas como $R_1 = 2R_2 = 4R_3 = \dots = 2^{n-1} R_n$ (Fig. 7.9.a) es complicado de implementar. Por eso en la práctica se utiliza la red R-2R cuyo comportamiento en cuanto a tensión de salida es el mismo (Fig. 7.9.b).

7.3.4 Conversores analógico-digitales

Un convertidor A/D, también llamado ADC, constituye el núcleo central de un sistema de adquisición de datos. Su función es la de transformar una señal continuamente variable en el tiempo en una sucesión unívoca de unos y ceros, es decir, en información binaria. Usualmente será necesario acondicionar la señal de entrada al ADC, bien sea atenuando, bien sea amplificando. En ocasiones, debido a la naturaleza de las señales a digitalizar, deberán utilizarse circuitos de muestreo especiales del tipo 'sample and hold'.

Veremos el fundamento de los tres tipos de conversores A/D utilizados más ampliamente y que son:

- 1) Tipo paralelo
- 2) Tipo contador y de aproximaciones sucesivas
- 3) Tipo integrador de doble rampa.

1) Conversor A/D tipo paralelo

Es el más sencillo de comprender ya que es simplemente una red de comparadores. Su funcionamiento no es secuencial sino que realiza simultáneamente 2^n comparaciones entre la señal V y 2^n niveles predeterminados. El resultado de estas comparaciones son 2^n señales digitales que son codificadas mediante un circuito combinacional. En la figura 7.10 vemos su diagrama de bloques funcionales.

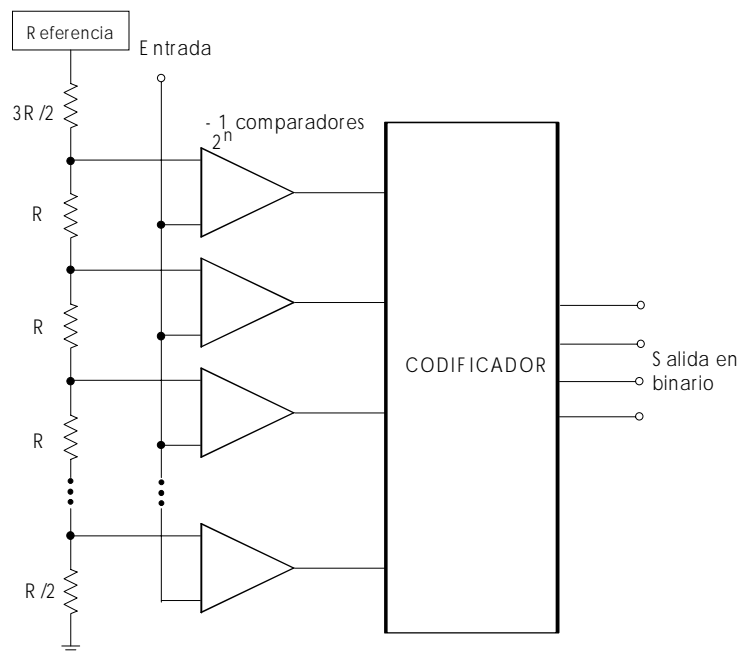


Fig. 7.10 Conversor A/D de tipo paralelo

La velocidad de este circuito puede ser muy alta ya que sólo está limitada por la del conjunto de comparadores del circuito lógico. Con este tipo se puede llegar a ritmos de conversión por encima de los 100MHz para 8 bits. Su principal inconveniente es el gran número de comparadores que requiere, lo que limita el número de bits con el que trabaja. Este es el tipo de convertidor más rápido, pero tiene el inconveniente de que su complejidad crece con el número de bits. En la práctica se construye para cuatro, seis u ocho bits como máximo (16, 64 ó 256 comparadores respectivamente). El precio está relacionado con esta complejidad.

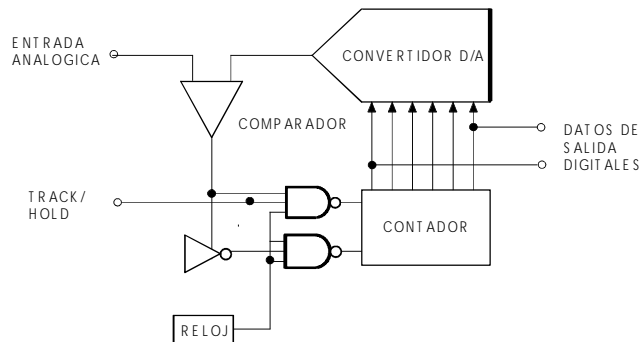


Fig. 7.11 Conversor A/D de tipo contador

2) Conversor A/D tipo contador y de aproximaciones sucesivas

Estos dos tipos están relacionados ya que ambos usan conversores D/A cuya salida comparan con la señal de entrada para obtener un valor digital de ella.

El de tipo contador es uno de los más simples y baratos. En la figura 7.11 vemos su diagrama en bloques funcionales.

Al comienzo de la conversión se permite el paso de los impulsos del reloj a la entrada del contador con lo que éste comienza a contar aquellos. A medida que el contador avanza cambia la salida del conversor D/A en escalera y esta salida se compara con la tensión analógica de entrada. Cuando llegamos a la igualdad el comparador cambia de estado y su salida bloquea la entrada de impulsos al contador. En este momento se ha acabado la conversión y el resultado digital de salida está contenido en las salidas del contador.

Este conversor tiene como ventajas la simplicidad, el bajo costo y su buena precisión y como gran desventaja su baja velocidad.

El conversor de aproximaciones sucesivas es probablemente el de uso más generalizado debido a que combina gran resolución y gran velocidad. En estos conversores se opera con un tiempo de conversión fijo por bit e independiente del valor de la entrada analógica. Este método se ilustra en la figura 7.12 y opera por comparaciones sucesivas de la tensión analógica de entrada con la salida del conversor D/A bit a bit.

Al comenzar el ciclo de conversión el bit más significativo (MSB) del conversor D/A (que es 1/2 del fondo de escala) aparece en su salida y es comparado con la entrada. Si es menor que ésta, se deja metido este bit y se intenta la misma operación con el bit siguiente. Si el MSB es mayor que la entrada éste es rechazado antes de pasar a meter el siguiente bit. Este proceso se continua de esta forma hasta el bit menos significativo (LSB), después del cual en el contador de salida tenemos el número digital correspondiente. Este contador constituye en este instante el registro de salida.

Con este método podemos conseguir velocidades tan altas como 100 nanosegundos por bit, además este tipo de conversores es bastante preciso y pueden trabajar en doble polaridad restando de la entrada una corriente o tensión equivalente a 1/2 FS.

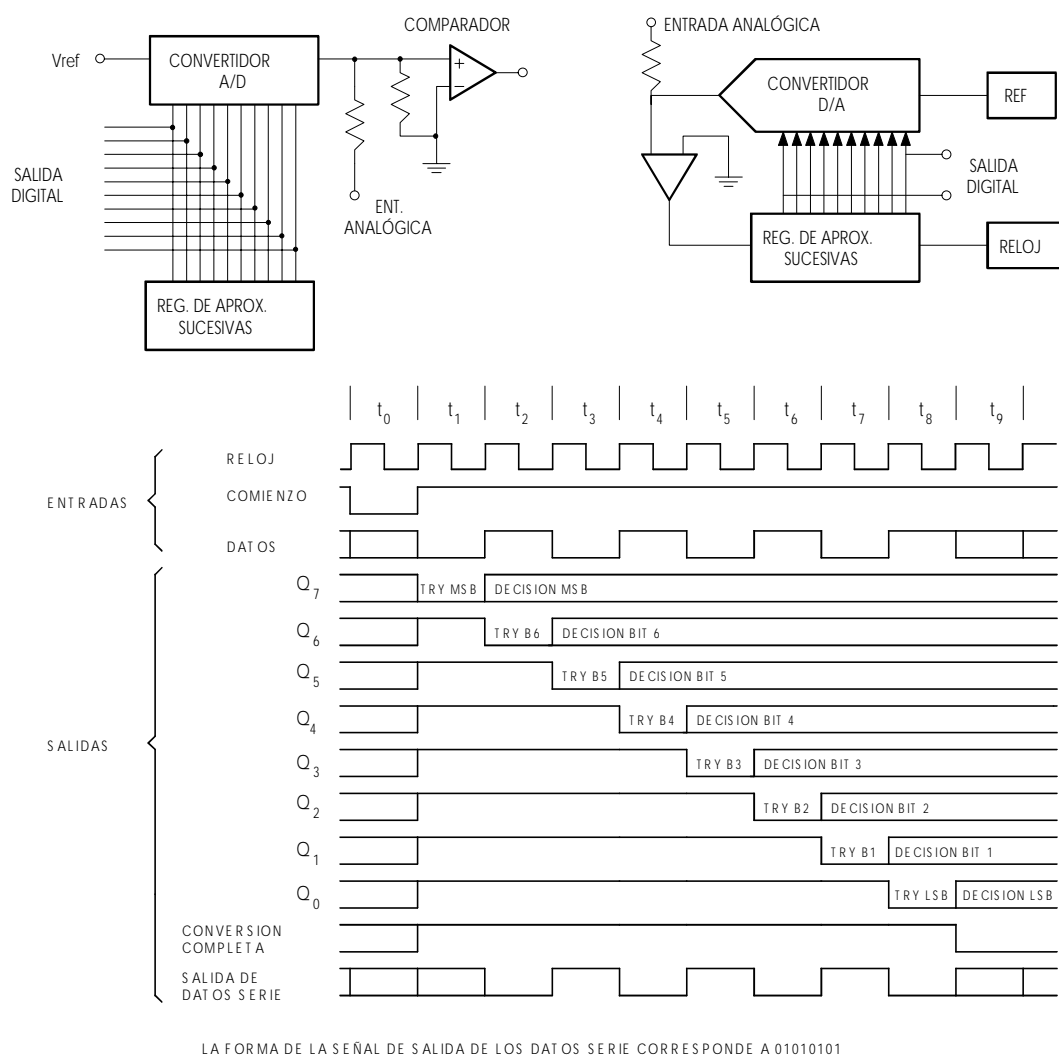


Fig. 7.12 Conversor A/D de aproximaciones sucesivas

3) Conversor A/D tipo integrador de doble rampa

Los conversores A/D de tipo integrador operan por el método indirecto de convertir un voltaje en un periodo de tiempo que posteriormente es medido por un contador. Hay muchos conversores usando este principio, pero el más popular y de más amplia utilización es el de doble rampa. Su diagrama funcional se representa en la figura 7.13.

La conversión comienza al conmutar la tensión desconocida de entrada a la entrada del integrador. Esto hace que el condensador comience a cargarse hasta una determinada tensión. La pendiente de carga será tanto más elevada cuanto mayor sea la tensión de entrada. Este proceso de carga se lleva a cabo durante un número fijo de pulsos al cabo del cual, el condensador tendrá un determinado nivel de carga que depende exclusivamente de la tensión desconocida de entrada. Una vez pasada esta primera fase, se conecta la entrada a la tensión de referencia (se desconecta la entrada) y el condensador comenzará a descargarse hasta alcanzar esta tensión de referencia. El tiempo que tarda en producirse esta descarga, es proporcional a la tensión inicial a la que se cargó durante la primera fase y que dependía únicamente de la tensión desconocida de entrada. Este tiempo es directamente proporcional a la tensión de entrada.

El integrador generará una rampa descendente que cruzará por el nivel de disparo del comparador en cuyo momento el contador es parado. La tensión de salida es:

$$E_{IN} = \frac{T_2}{T_1} V_{REF}$$

donde T_1 y T_2 son los números de cuentas acumuladas en el contador de estos intervalos y se obtendrán directamente de lo registrado en el contador al final del proceso.

El método de doble rampa posee ciertas ventajas. La precisión es independiente de la frecuencia del reloj y del valor de la capacidad de integración siempre que sean estables durante un periodo de conversión y sólo depende de la precisión y estabilidad de la referencia.

La resolución está limitada básicamente sólo por la del comparador. Además este conversor da un excelente rechazo al ruido por ser de tipo integrador. La principal desventaja de este método es que el tiempo de conversión es relativamente largo.

La figura 7.14 compara los productos comerciales que usan estos tres métodos en términos de coste frente a velocidad.

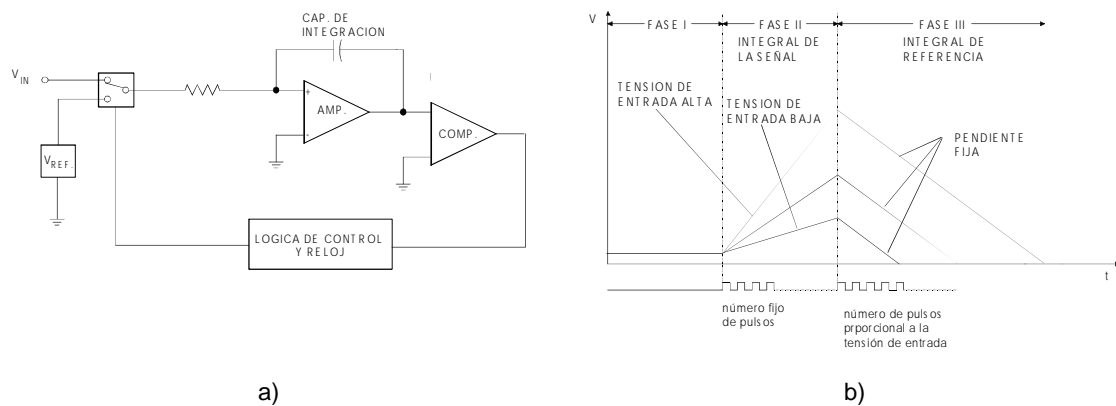


Fig. 7.13 a) Convertor integrador de doble rampa, b) la conversión se produce en tres fases distintas

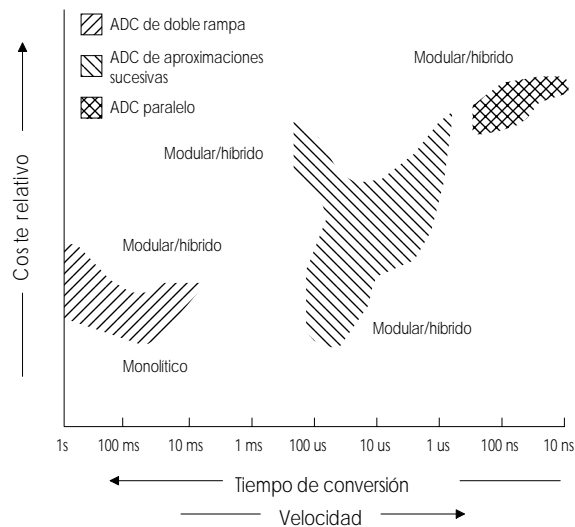


Fig. 7.14 Comparaciones entre diferentes convertidores A/D en términos de coste de velocidad

7.3.5 Multiplexores analógicos

Estos circuitos se usan para compartir el tiempo a la entrada de un convertidor A/D entre varios canales analógicos de información. Son útiles para evitar tener que disponer varios

convertidores A/D. Los tipos más usuales son de 4, 8 y 16 canales conectados en forma simple o diferencial. Un multiplexor analógico consta de un grupo de interruptores analógicos ordenados con entradas conectadas a los canales analógicos individuales y una salida común como se muestra en la figura 7.15. Los interruptores se pueden direccionar con un código digital de entrada. Se usan generalmente interruptores 'MOSFET' los cuales se pueden conectar directamente a la carga de salida si ésta tiene una impedancia suficientemente alta o, en su defecto, necesitaremos usar un amplificador 'buffer' de salida. Podemos conseguir que la impedancia de entrada del 'buffer' sea del orden de $10^9 \Omega$ en cuyo caso el error de transferencia debido a la resistencia del interruptor es despreciable, ya que ésta suele ser del orden de los $2K\Omega$. En la figura vemos un circuito multiplexor de 8 canales.

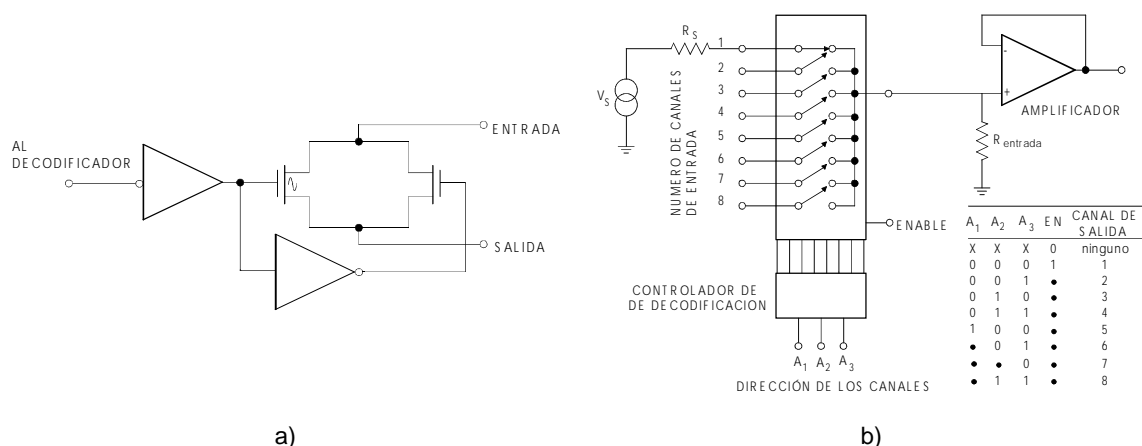


Fig. 7.15 a) Interruptor analógico basado en transistores MOS. b) Ejemplo de multiplexor analógico

7.3.6 Circuitos de muestreo y retención

Los circuitos de muestreo y retención o S&H (sample and hold') se usan ampliamente en el procesamiento de señales analógicas y en sistemas de conversión de datos para almacenar de forma precisa, una tensión analógica durante un tiempo que puede estar comprendido entre algo menos de $1\mu\text{seg.}$ y varios minutos. Esta característica les da importancia en aplicaciones que incluyen sistemas de adquisición de datos, sistemas simultáneos de muestreo y retención, en los convertidores A/D, osciloscopios de muestreo, multímetros digitales, filtros reconstructores de señal y circuitos analógicos de computación. Este tipo de circuitos son necesarios, para que el conversor A/D de un sistema de adquisición de datos disponga de una señal estable a su entrada durante el periodo de conversión.

Estos circuitos cumplen la misión de muestreo y retención que se vieron en la teoría de muestreo. Aquí discutiremos brevemente la configuración del circuito. Los circuitos de muestreo y retención se usan junto con los conversores A/D ó D/A. Con los conversores A/D se usan para acortar el tiempo de apertura para el conversor, al muestrear rápidamente la señal y después mantener su valor hasta que la conversión finalice. En los conversores D/A para mantener la salida un tiempo mayor.

Un circuito 'sample and hold' (Fig. 7.16) está formado básicamente por un interruptor y un condensador. Cuando el interruptor está cerrado el circuito está en el modo de muestreo ('sampling mode') y seguirá a una señal variable de entrada. Cuando el interruptor se abre el circuito está en el modo de mantenimiento ('hold mode') y retiene una tensión en el condensador durante cierto tiempo que depende de éste y de las fugas del interruptor.

Los circuitos de muestreo y retención ('sample and hold') prácticos también usan amplificadores 'buffer' de entrada y salida, y sofisticadas técnicas de conmutación.

El amplificador 'buffer' de salida debe tener un 'FET' (transistor de efecto campo) de baja corriente de entrada (alta impedancia) para que el efecto de fugas del condensador sea lo más pequeño posible.

En la figura 7.16 vemos varias configuraciones de circuitos 'sample and hold' usadas normalmente. Algunos se usan para circuitos de muestreo y retención rápidos. Otro es una configuración en lazo cerrado con un integrador operacional en la línea de realimentación del 'buffer' de entrada. Este circuito tiene gran precisión y linealidad.

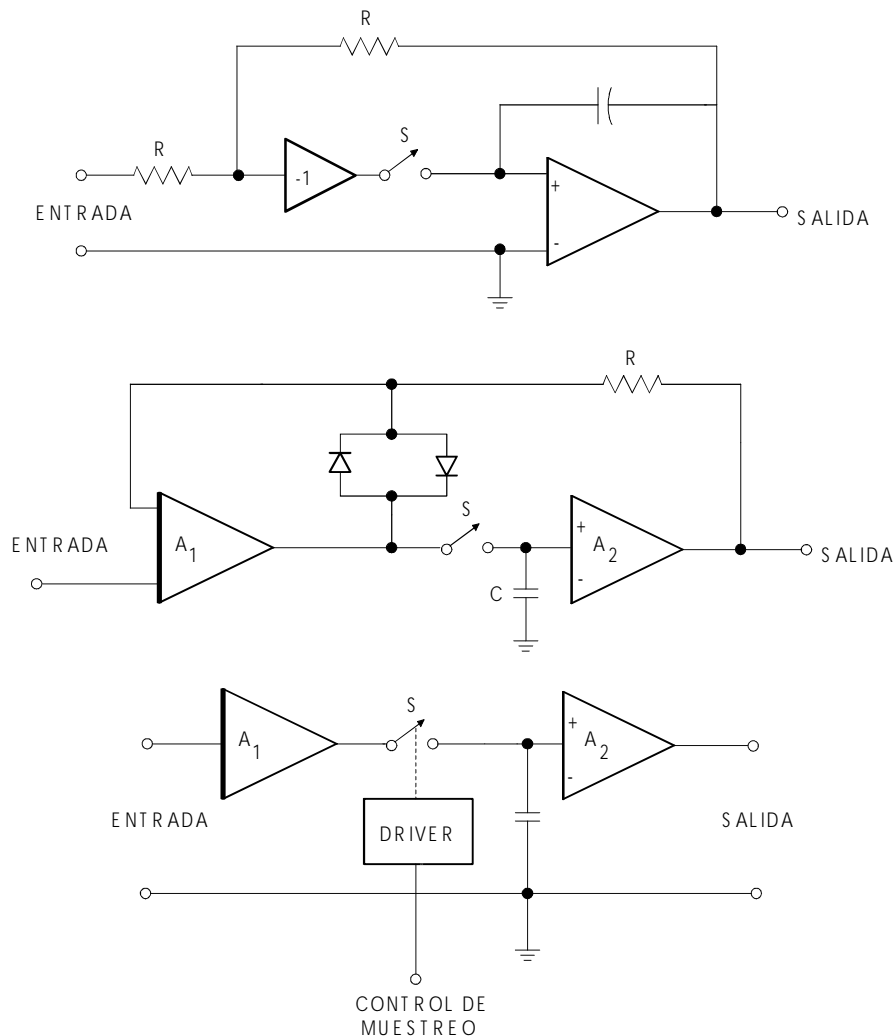


Fig. 7.16 Circuitos de muestreo y retención "sample & hold"

7.3.7 Modos de conexión de un sistema de adquisición de datos a un ordenador

Existe básicamente cuatro modos de conexión de un convertidor analógico-digital a un ordenador:

1) Adquisición del valor más reciente

En este método el convertidor está funcionando continuamente y al final de cada conversión el registro de salida es actualizado con el nuevo valor. El ordenador procede simplemente a leer

este registro en el momento que lo necesita. Este registro es actualizado a la velocidad máxima de actuación del convertidor.

2) Comienza y espera

En este método el ordenador controla el proceso poniendo en marcha el convertidor cuando lo necesita y espera la señal de (EOC, 'End Of Conversión') de fin de la conversión que le indica que ésta ha concluido y que en el registro de salida del convertidor se encuentra el valor deseado, a continuación el ordenador lee este valor. Otra técnica consiste en esperar un tiempo superior al de conversión y leer entonces el registro de salida del convertidor. Este procedimiento es bastante sencillo de implementar pero el ordenador no puede hacer otra cosa mientras espera a la conversión.

3) Utilizar interrupciones

Este método hace uso de las capacidades de interrupción del ordenador. Bien por un reloj o por el propio ordenador se da la orden de inicio de la conversión y el ordenador sigue haciendo otro programa. Cuando el convertidor termina (la señal EOC) produce una interrupción al ordenador, obligándole a abandonar la tarea que está realizando para atender a la rutina de servicio de la toma de datos. A continuación prosigue su tarea.

4) Utilizar acceso directo a memoria (DMA)

El acceso directo a memoria es la manera más eficaz de transferir datos a alta velocidad. Este método permite la transferencia Entrada/Salida sin intervención de programa. La transferencia se efectúa a través de canales especiales que 'roban' ciclos del bus sin que el procesador intervenga. Sólo tienen sentido para la introducción de un bloque de datos, por lo que su utilidad en adquisición de datos se reduce al estudio de transitorios o a transferencias de datos entre equipos que posean memoria propia y la del ordenador.

7.3.8 Especificaciones y parámetros característicos

ESPECIFICACIONES DEL SISTEMA DE ENTRADA Y SALIDAS ANALÓGICAS

ENTRADAS	SALIDAS
Número máximo de señales	Número máximo de señales
Frecuencia máxima de lectura	Tiempo de conversión
Márgenes de tensión	Margen de la tensión
Tensión máxima accidental	Impedancia de salida/carga
Tipo de señal y masa	Precisión
Impedancia de salida del transductor	Estabilidad
Impedancia de entrada en el ordenador	
Desequilibrio de impedancias en el sistema	
Ruido en modo común	
Espectro del ruido normal	
Precisión	
Estabilidad	

ESPECIFICACIONES DEL SISTEMA DE ENTRADA Y SALIDAS DIGITALES

ENTRADAS	SALIDAS
Número máximo de señales	Número máximo de señales
Frecuencia máxima	Frecuencia máxima
Sincronización	Sincronización
Duración mínima de la señal antes de su lectura	Duración de la señal
Duración máxima o mínima de la señal	Impedancia salida/carga
Impedancia máxima en circuito cerrado	Límites de tensión y corriente
Impedancia mínima circuito abierto	
Límites de tensión y corriente	
Tensión máxima accidental	
Tipo de señal y masa	

ESPECIFICACIONES PARA EL CABLEADO INSTRUMENTACIÓN-ORDENADOR

Unifilar / bifilar
 Trenzado / paralelo
 Pantalla: tipo, recubrimiento, masa, etc.
 Sección
 Aislamiento y protección
 Número máximo de cables juntos
 Separación de cables con otro tipo de señal (alterna, potencia, etc.)
 Longitud máxima de cable permitida
 Tipo de conexión terminal