CAPÍTULO 4

Memoria caché

PUNTOS CLAVE

- La memoria de un computador tiene una organización jerárquica. En el nivel superior (el más próximo al procesador) están los registros del procesador. A continuación se encuentran uno o más niveles de caché, denominados L1, L2, etc. Posteriormente la memoria principal, normalmente construida con memorias dinámicas de acceso aleatorio (DRAM). Todas ellas se consideran memorias internas del computador. La jerarquía prosigue con la memoria externa, siendo el siguiente nivel usualmente un disco duro fijo, y uno o más niveles de soportes extraíbles tales como discos ópticos y cintas magnéticas.
- A medida que descendemos en la jerarquía de memoria disminuye el coste por bit, aumenta la capacidad y crece el tiempo de acceso. Sería deseable poder utilizar solo la memoria más rápida, pero al ser la más costosa se llega a un compromiso entre tiempo de acceso y coste, empleando más cantidad de memoria más lenta. La estrategia a seguir consiste en organizar los datos y los programas en memoria de manera que las palabras de memoria necesarias estén normalmente en la memoria más rápida.
- En general, es probable que la mayoría de los accesos futuros a la memoria principal, por parte del procesador, sean a posiciones accedidas recientemente. Por eso la caché automáticamente retiene una copia de algunas de las palabras de la DRAM utilizadas recientemente. Si la caché se diseña adecuadamente, la mayor parte del tiempo el procesador solicitará palabras de memoria que están ya en la caché.

Las memorias de los computadores, aunque parezcan conceptualmente sencillas, presentan tal vez la más amplia diversidad de tipos, tecnología, estructura, prestaciones y coste, de entre todos los componentes de un computador. Ninguna tecnología es óptima para satisfacer las necesidades de memoria de un computador. En consecuencia, un computador convencional está equipado con una jerarquía de subsistemas de memoria, algunos internos (directamente accesibles por el procesador), y otros externos (accesibles por el procesador mediante módulos de entrada/salida).

Este capítulo y el siguiente se centran en el estudio de la memoria interna, mientras que el Capítulo 6 se dedicará a la memoria externa. Para comenzar, en la primera sección de este capítulo examinaremos características clave de las memorias de un computador. El resto del capítulo se dedica al estudio de un elemento esencial de cualquier computador moderno: la memoria caché.

4.1. CONCEPTOS BÁSICOS SOBRE SISTEMAS DE MEMORIA DE COMPUTADORES

CARACTERÍSTICAS DE LOS SISTEMAS DE MEMORIA

El complejo tema de las memorias es más abordable si clasificamos los sistemas de memoria según sus características clave. Las más importantes se listan en la Tabla 4.1.

Tabla 4.1. Características clave de los sistemas de memoria de computadores.

Ubicación	**Pestaciones**
Procesador	Tiempo de acceso
Interna (principal)	Tiempo de ciclo
Externa (secundaria)	Velocidad de transferencia
Capacidad	**Dispositivo físico**
Tamaño de la palabra	Semiconductor
Número de palabras	Soporte magnético
Unidad de transferencia	Soporte óptico
Palabra	Magneto-óptico
Bloque	**Características físicas**
Método de acceso	Volátil/no volátil
Acceso secuencial	Borrable/no borrable
Acceso directo	**Organización**
Acceso aleatorio	
Acceso asociativo	

El término **ubicación** que aparece en la Tabla 4.1 indica si la memoria es interna o externa al computador. La memoria interna suele identificarse con la memoria principal. Sin embargo hay además otras formas de memoria interna. El procesador necesita su propia memoria local en forma de registros (*véase* por ejemplo la Figura 2.3). Además, como veremos, la unidad de control del procesador también puede necesitar su propia memoria interna. Postponemos la discusión de estos dos últimos tipos de memoria interna para capítulos posteriores. La memoria caché es también otro tipo de memoria interna. La memoria externa consta de dispositivos periféricos de almacenamiento, tales como discos y cintas, que son accesibles por el procesador a través de controladores de E/S.

Una característica obvia de las memorias es su **capacidad**. Para memorias internas se expresa normalmente en términos de bytes (1 byte = 8 bits) o de palabras. Longitudes de palabra comunes son 8, 16, y 32 bits. La capacidad de las memorias externas se suele expresar en bytes.

Un concepto relacionado es la **unidad de trasferencia**. Para memorias internas, la unidad de transferencia es igual al número de líneas de entrada/salida de datos del módulo de memoria. A menudo es igual a la longitud de palabra, pero suele ser mayor, por ejemplo 64, 128, o 256 bits. Para aclararlo consideremos tres conceptos relacionados con la memoria interna:

- **Palabra:** es la unidad «natural» de organización de la memoria. El tamaño de la palabra suele coincidir con el número de bits utilizados para representar números y con la longitud de las instrucciones. Por desgracia hay muchas excepciones. Por ejemplo, el CRAY C90 tiene una longitud de palabra de 64 bits, pero utiliza una representación de números enteros de 46 bits. El VAX tiene una gran variedad de longitudes de instrucción, expresadas como múltiplos de bytes, y una longitud de palabra de 32 bits.
- **Unidades direccionables:** en algunos sistemas la unidad direccionable es la palabra. Sin embargo muchos de ellos permiten direccionar a nivel de bytes. En cualquier caso, la relación entre la longitud A de una dirección y el número N de unidades direccionables, es $2^A = N$.

- **Unidad de transferencia:** para la memoria principal es el número de bits que se leen o escriben en memoria a la vez. La unidad de transferencia no tiene por qué coincidir con una palabra o con una unidad direccionable. Para la memoria externa, los datos se transfieren normalmente en unidades más grandes que la palabra denominadas bloques.

Otro distintivo entre tipos de memorias es el **método de acceso**, que incluye las siguientes variantes:

- **Acceso secuencial:** la memoria se organiza en unidades de datos llamadas registros. El acceso debe realizarse con una secuencia lineal específica. Se hace uso de información almacenada de direccionamiento que permite separar los registros y ayudar en el proceso de recuperación de datos. Se utiliza un mecanismo de lectura/escritura compartida que debe ir trasladándose desde su posición actual a la deseada, pasando y obviando cada registro intermedio. Así pues, el tiempo necesario para acceder a un registro dado es muy variable. Las unidades de cinta que se tratan en el Capítulo 6 son de acceso secuencial.
- **Acceso directo:** como en el caso de acceso secuencial, el directo tiene asociado un mecanismo de lectura/escritura. Sin embargo, los bloques individuales o registros tienen una dirección única basada en su dirección física. El acceso se lleva a cabo mediante un acceso directo a una vecindad dada, seguido de una búsqueda secuencial, bien contando, o bien esperando hasta alcanzar la posición final. De nuevo el tiempo de acceso es variable. Las unidades de disco, que se tratan en el Capítulo 6, son de acceso directo.
- **Acceso aleatorio (*random*):** cada posición direccionable de memoria tiene un único mecanismo de acceso cableado físicamente. El tiempo para acceder a una posición dada es constante e independiente de la secuencia de accesos previos. Por tanto, cualquier posición puede seleccionarse «aleatoriamente» y ser direccionada y accedida directamente. La memoria principal y algunos sistemas de caché son de acceso aleatorio.
- **Asociativa:** es una memoria del tipo de acceso aleatorio que permite hacer una comparación de ciertas posiciones de bits dentro de una palabra buscando que coincidan con unos valores dados, y hacer esto para todas las palabras simultáneamente. Una palabra es por tanto recuperada basándose en una porción de su contenido en lugar de su dirección. Como en las memorias de acceso aleatorio convencionales, cada posición tiene su propio mecanismo de direccionamiento, y el tiempo de recuperación de un dato es una constante independiente de la posición o de los patrones de acceso anteriores. Las memorias caché pueden emplear acceso asociativo.

Desde el punto de vista del usuario, las dos características más importantes de una memoria son su capacidad y sus **prestaciones**. Se utilizan tres parámetros de medida de prestaciones:

- **Tiempo de acceso (latencia):** para memorias de acceso aleatorio es el tiempo que tarda en realizarse una operación de escritura o de lectura, es decir, el tiempo que transcurre desde el instante en el que se presenta una dirección a la memoria hasta que el dato, o ha sido memorizado, o está disponible para su uso. Para memorias de otro tipo, el tiempo de acceso es el que se tarda en situar el mecanismo de lectura/escritura en la posición deseada.
- **Tiempo de ciclo de memoria:** este concepto se aplica principalmente a las memorias de acceso aleatorio y consiste en el tiempo de acceso y algún tiempo más que se requiere antes de que pueda iniciarse un segundo acceso a memoria. Este tiempo adicional puede que sea necesario

para que finalicen las transiciones en las líneas de señal o para regenerar los datos en el caso de lecturas destructivas. Tenga en cuenta que el tiempo de ciclo de memoria depende de las características del bus del sistema y no del procesador.

- **Velocidad de transferencia:** es la velocidad a la que se pueden transferir datos a, o desde, una unidad de memoria. Para memorias de acceso aleatorio coincide con el inverso del tiempo de ciclo.

Para otras memorias se utiliza la siguiente relación:

$$T_N = T_A + \frac{N}{R}$$

donde:

T_N = Tiempo medio de escritura o de lectura de N bits

T_A = Tiempo de acceso medio

N = Número de bits

R = Velocidad de transferencia, en bits por segundo (bps)

Se han empleado **soportes físicos** muy diversos para las memorias. Las más comunes en la actualidad son las memorias semiconductoras, las memorias de superficie magnética, utilizadas para discos y cintas, y las memorias ópticas y magneto-ópticas.

Del almacenamiento de datos son importantes varias **características físicas**. En memorias volátiles la información se va perdiendo o desaparece cuando se desconecta la alimentación. En las memorias no volátiles la información, una vez grabada, permanece sin deteriorarse hasta que se modifique intencionadamente; no se necesita la fuente de alimentación para retener la información. Las memorias de superficie magnética son no volátiles. Las memorias semiconductoras pueden ser volátiles o no volátiles. Las memorias no borrables no pueden modificarse, salvo que se destruya la unidad de almacenamiento. Las memorias semiconductoras de este tipo se conocen por el nombre de *memorias de solo lectura* (ROM, *Read Only Memory*). Una memoria no borrable es necesariamente no volátil.

En memorias de acceso aleatorio, su **organización** es un aspecto clave de diseño. Por *organización* se entiende su disposición o estructura física en bits para formar palabras. Como explicaremos pronto, la estructura más obvia no es siempre la utilizada en la práctica.

JERARQUÍA DE MEMORIA

Las restricciones de diseño de la memoria de un computador se pueden resumir en tres cuestiones: ¿cuánta capacidad? ¿cómo de rápida? ¿de qué coste?

La cuestión del tamaño es un tema siempre abierto. Si se consigue hasta una cierta capacidad, probablemente se desarrollarán aplicaciones que la utilicen. La cuestión de la rapidez es, en cierto sentido, fácil de responder. Para conseguir las prestaciones óptimas, la memoria debe seguir al procesador. Es decir, cuando el procesador ejecuta instrucciones, no es deseable que tenga que detenerse a la espera de instrucciones o de operandos. La última de las cuestiones anteriores también debe tenerse en cuenta. En la práctica, el coste de la memoria debe ser razonable con relación a los otros componentes.

Como es de esperar, existe un compromiso entre las tres características clave de coste, capacidad, y tiempo de acceso. En un momento dado, se emplean diversas tecnologías para realizar los sistemas de memoria. En todo el espectro de posibles tecnologías se cumplen las siguientes relaciones:

- A menor tiempo de acceso, mayor coste por bit.
- A mayor capacidad, menor coste por bit.
- A mayor capacidad, mayor tiempo de acceso.

El dilema con que se enfrenta el diseñador está claro. El diseñador desearía utilizar tecnologías de memoria que proporcionen gran capacidad, tanto porque esta es necesaria como porque el coste por bit es bajo. Sin embargo, para satisfacer las prestaciones requeridas, el diseñador necesita utilizar memorias costosas, de capacidad relativamente baja y con tiempos de acceso reducidos.

La respuesta a este dilema es no contar con un solo componente de memoria, sino emplear una **jerarquía de memoria**. La Figura 4.1 ilustra una jerarquía típica. Cuando se desciende en la jerarquía ocurre:

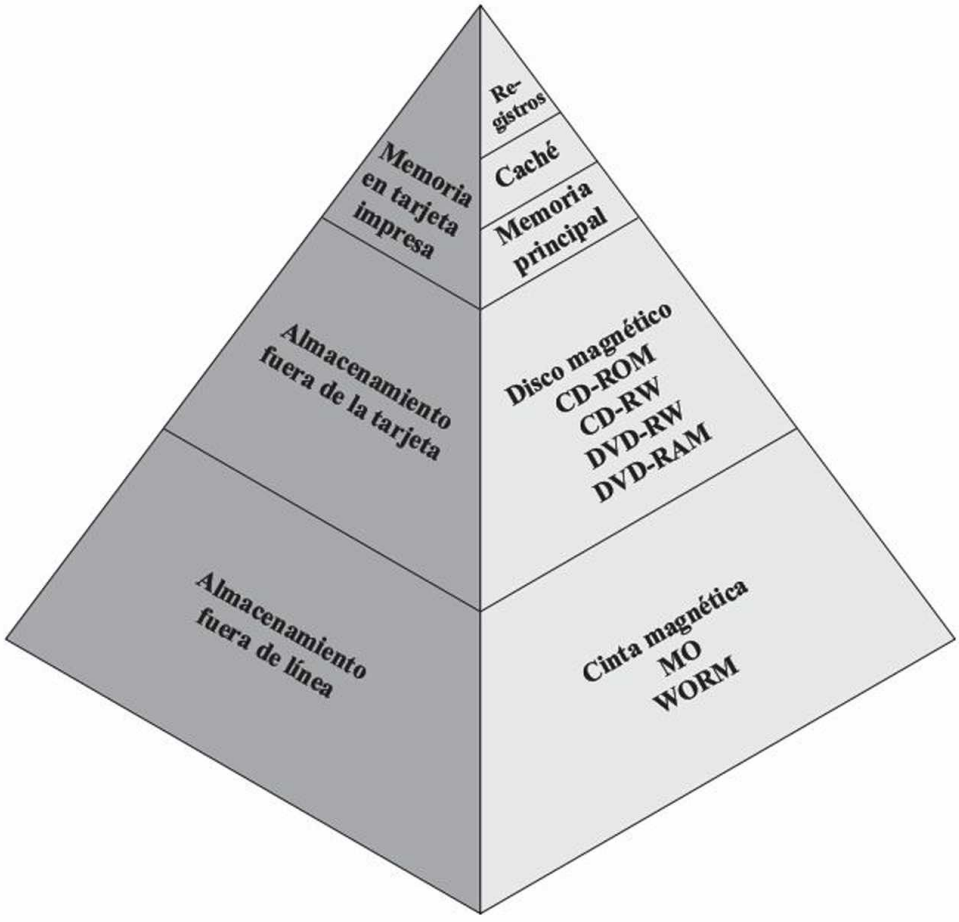

Figura 4.1. Jerarquía de memoria.

a) Disminuye el coste por bit.

b) Aumenta la capacidad.

c) Aumenta el tiempo de acceso.

d) Disminuye la frecuencia de accesos a la memoria por parte del procesador.

Así pues, memorias más pequeñas, más costosas y más rápidas, se complementan con otras más grandes, más económicas y más lentas. La clave del éxito de esta organización está en el último *item* (d): la disminución de la frecuencia de acceso. Examinaremos con detalle este concepto cuando hablemos de la caché (después, en este mismo capítulo) y de la memoria virtual (en el Capítulo 8), pero daremos aquí una breve explicación.

Ejemplo 4.1. Supongamos que el procesador tiene que acceder a dos niveles de la memoria. El nivel 1 contiene 1 000 palabras y tiene un tiempo de acceso de 0,01 μs. El nivel 2 contiene 100 000 palabras y tiene un tiempo de acceso de 0,1 μs. Supongamos que si la palabra a la que se va a acceder está en el nivel 1, el procesador accede a ella directamente. Si está en el nivel 2, entonces es primeramente transferida al nivel 1 y después accedida por el procesador. Por simplicidad ignoramos el tiempo necesario para que el procesador determine si la palabra está en un nivel u otro. La Figura 4.2 muestra la forma que en general tiene la curva que representa esta situación. La figura muestra el tiempo de acceso medio a una memoria de dos niveles, en función de la tasa de acierto H, donde H se define como la fracción del total de accesos a memoria encontrados en la memoria más rápida (por ejemplo, en la caché); T_1 es el tiempo de acceso al nivel 1, y T_2 el tiempo de acceso al nivel 2[1]. Como puede verse, para porcentajes altos de accesos al nivel 1, el tiempo de acceso total promedio es mucho más próximo al del nivel 1 que al del nivel 2.

En nuestro ejemplo, si suponemos que el 95 por ciento de los accesos a memoria se encuentran con éxito en la caché, entonces el tiempo medio para acceder a una palabra puede expresarse en la forma:

$$(0{,}95)\ (0{,}01\ \mu s) + (0{,}05)\ (0{,}01\ \mu s + 0{,}1\ \mu s) = 0{,}0095\ \mu s + 0{,}0055\ \mu s = 0{,}015\ \mu s$$

Como era deseable, el tiempo de acceso medio está mucho más próximo a 0,01 μs que a 0,1 μs.

En principio, el uso de dos niveles de memoria para reducir el tiempo de acceso medio funciona, pero solo si se aplican las condiciones (a) a (d) anteriores. Empleando diversas tecnologías se tiene todo un espectro de sistemas de memoria que satisfacen las condiciones (a) a (c). Afortunadamente, la condición (d) es también generalmente valida.

La base para la validez de la condición (d) es el principio conocido como **localidad de las referencias** [DENN68]. En el curso de la ejecución de un programa, las referencias a memoria por parte del procesador, tanto para instrucciones como para datos, tienden a estar agrupadas. Los programas normalmente contienen un número de bucles iterativos y subrutinas. Cada vez que se entra en un bucle o una subrutina, hay repetidas referencias a un pequeño conjunto de instrucciones. De manera similar, las

[1] Si la palabra accedida se encontraba en la memoria más rápida, se dice que se ha producido un **acierto**. Y si no se encontraba en la memoria más rápida, se dice que ha tenido lugar un **fallo.**

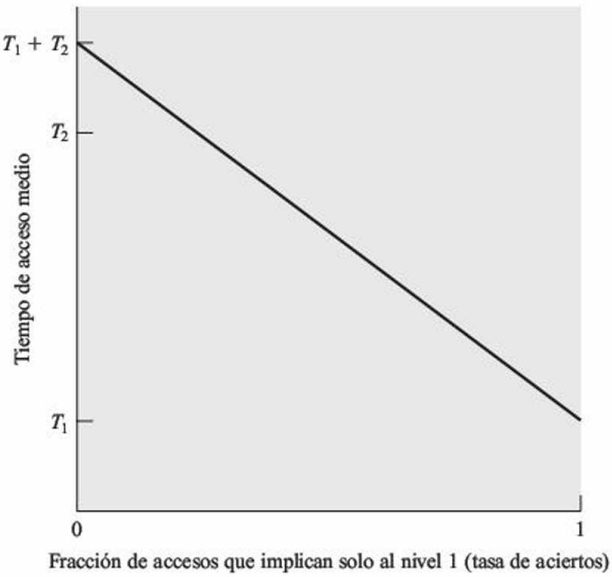

Figura 4.2. Prestaciones de una memoria de dos niveles sencilla.

operaciones con tablas o con matrices conllevan accesos a un conjunto de palabras de datos agrupadas. En periodos de tiempo largos, las agrupaciones (*clusters*) en uso cambian, pero en periodos de tiempo cortos, el procesador trabaja principalmente con agrupaciones fijas de referencias a memoria.

De acuerdo con lo anterior, es posible organizar los datos a través de la jerarquía de tal manera que el porcentaje de accesos a cada nivel siguiente más bajo sea sustancialmente menor que al nivel anterior. Considérese el ejemplo de dos niveles ya presentado, y que la memoria del nivel 2 contiene todos los datos e instrucciones de programa. Las agrupaciones actuales pueden ubicarse temporalmente en el nivel 1. De vez en cuando, una de las agrupaciones del nivel 1 tendrá que ser devuelta al nivel 2 a fin de que deje sitio para que entre otra nueva agrupación al nivel 1. En general, sin embargo, la mayoría de las referencias serán a instrucciones y datos contenidos en el nivel 1.

Este principio puede aplicarse a través de más de dos niveles de memoria, como sugiere la jerarquía mostrada en la Figura 4.1. El tipo de memoria más rápida, pequeña y costosa, lo constituyen los registros internos al procesador. Un procesador suele contener unas cuantas docenas de tales registros, aunque algunas máquinas contienen cientos de ellos. Descendiendo dos niveles, la memoria principal es el principal sistema de memoria interna del computador. Cada posición de memoria principal tiene una única dirección. La memoria principal es normalmente ampliada con una caché, que es más pequeña y rápida. La caché no suele estar visible al programador, y realmente tampoco al procesador. Es un dispositivo para escalonar las transferencias de datos entre memoria principal y los registros del procesador a fin de mejorar las prestaciones.

Las tres formas de memoria que acabamos de describir son, normalmente, volátiles y de tecnología semiconductora. El uso de tres niveles aprovecha la variedad existente de tipos de memorias semiconductoras, que difieren en velocidad y coste. El almacenamiento de datos de forma más permanente se hace en dispositivos de memoria masiva, de los cuales los más comunes son el disco duro y los dispositivos extraíbles, tales como discos extraíbles, cintas y dispositivos ópticos de almacenamiento.

Las memorias externas no volátiles o permanentes se denominan también memorias secundarias o auxiliares. Se utilizan para almacenar programas y ficheros de datos, y suelen estar visibles al programador solo en términos de ficheros y registros, en lugar de bytes aislados o de palabras. El disco se emplea además para proporcionar una ampliación de la memoria principal conocida como memoria virtual, que será tratada en el Capítulo 8.

En la jerarquía pueden incluirse otras formas de memoria. Por ejemplo, los grandes computadores de IBM incluyen una forma de memoria interna conocida como almacenamiento extendido. Este utiliza una tecnología semiconductora que es más lenta y menos costosa que la de la memoria principal. Estrictamente hablando, esta memoria no encaja en la jerarquía sino que es una ramificación lateral: los datos pueden transferirse entre memoria principal y el almacenamiento extendido pero no entre este y la memoria externa. Otras formas de memoria secundaria incluyen los discos ópticos y los magneto-ópticos. Finalmente, mediante software se pueden añadir más niveles a la jerarquía. Una parte de la memoria principal puede utilizarse como almacén intermedio (*buffer*) para guardar temporalmente datos que van a ser volcados en disco. Esta técnica, a veces denominada caché de disco [2], mejora las prestaciones de dos maneras:

- Las escrituras en disco se hacen por grupos. En lugar de muchas transferencias cortas de datos, tenemos pocas transferencias largas. Esto mejora las prestaciones del disco y minimiza la participación del procesador.
- Algunos datos destinados a ser escritos como salidas pueden ser referenciados por un programa antes de que sean volcados en disco. En ese caso, los datos se recuperan rápidamente desde la caché software en lugar de hacerlo lentamente de disco.

El Apéndice 4A examina las implicaciones sobre prestaciones de las estructuras de memoria multinivel.

4.2. PRINCIPIOS BÁSICOS DE LAS MEMORIAS CACHÉ

El objetivo de la memoria caché es lograr que la velocidad de la memoria sea lo más rápida posible, consiguiendo al mismo tiempo un tamaño grande al precio de memorias semiconductoras menos costosas. El concepto se ilustra en la Figura 4.3. Hay una memoria principal relativamente grande y más

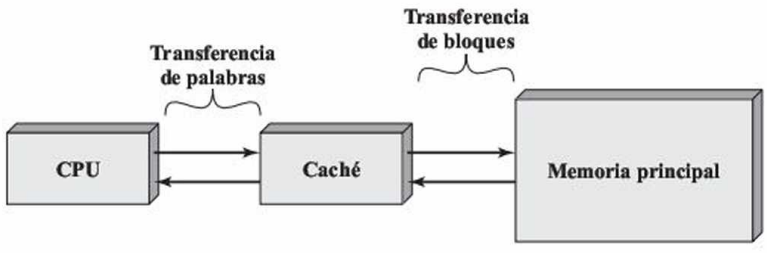

Figura 4.3. Memorias caché y principal.

[2] La caché de disco generalmente es una técnica software y no es estudiada en este libro. *Véase* [STAL05] para una discusión del tema.

lenta, junto con una memoria caché más pequeña y rápida. La caché contiene una copia de partes de la memoria principal. Cuando el procesador intenta leer una palabra de memoria, se hace una comprobación para determinar si la palabra está en la caché. Si es así, se entrega dicha palabra al procesador. Si no, un bloque de memoria principal, consistente en un cierto número de palabras, se transfiere a la caché y después la palabra es entregada al procesador. Debido al fenómeno de localidad de las referencias, cuando un bloque de datos es capturado por la caché para satisfacer una referencia a memoria simple, es probable que se hagan referencias futuras a la misma posición de memoria o a otras palabras del mismo bloque.

La Figura 4.4 describe la estructura de un sistema de memoria caché/principal. La memoria principal consta de hasta 2^n palabra direccionables, teniendo cada palabra una única dirección de n bits. Esta memoria la consideramos dividida en un número de bloques de longitud fija, de K palabras por bloque. Es decir, hay $M = 2^n/K$ bloques. La caché consta de C **líneas.** Cada línea contiene K palabras, más una etiqueta de unos cuantos bits; denominándose tamaño de línea al número de palabras que hay en la línea. El número de líneas es considerablemente menor que el número de bloques de memoria principal ($C << M$). En todo momento, un subconjunto de los bloques de memoria reside en líneas de la caché. Si se lee una palabra de un bloque de memoria, dicho bloque es transferido a una de las líneas de la caché. Ya que hay más bloques que líneas, una línea dada no puede dedicarse unívoca y permanentemente a un bloque. Por consiguiente, cada línea incluye una **etiqueta** que identifica qué bloque particular almacena. La etiqueta es usualmente una porción de la dirección de memoria principal, como describiremos más adelante en esta sección.

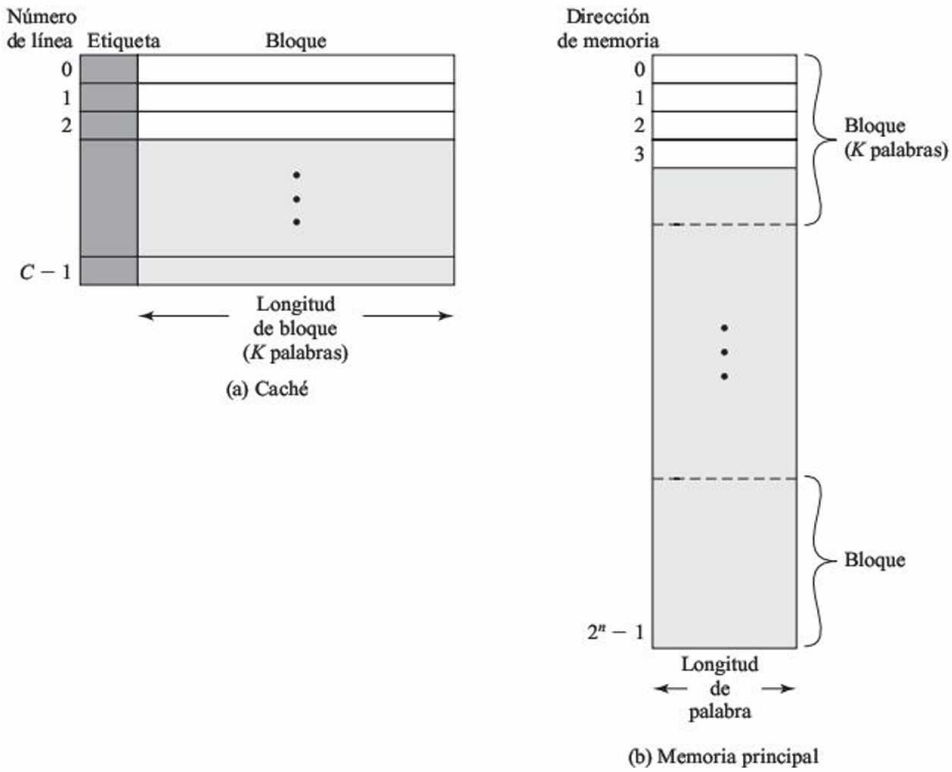

Figura 4.4. Estructura de memoria caché/principal.

La Figura 4.5 ilustra una operación de lectura. El procesador genera la dirección, RA, de una palabra a leer. Si la palabra está en la caché, es entregada al procesador. Si no, el bloque que contiene dicha palabra se carga en la caché, y la palabra después es llevada al procesador. La Figura 4.5 indica cómo estas dos últimas operaciones se realizan en paralelo y refleja la organización mostrada en la Figura 4.6, que es típica en las organizaciones de caché actuales. En ella, la caché conecta con el procesador mediante líneas de datos, de control y de direcciones. Las líneas de datos y de direcciones conectan también con buffers de datos y de direcciones que las comunican con un bus del sistema a través del cual se accede a la memoria principal. Cuando ocurre un acierto de caché, los buffers de datos y de direcciones se inhabilitan, y la comunicación tiene lugar solo entre procesador y caché, sin tráfico en el bus. Cuando ocurre un fallo de caché, la dirección deseada se carga en el bus del sistema y el dato es llevado, a través del buffer de datos, tanto a la caché como al procesador. En otras formas de organización, la caché se interpone físicamente entre el procesador y la memoria

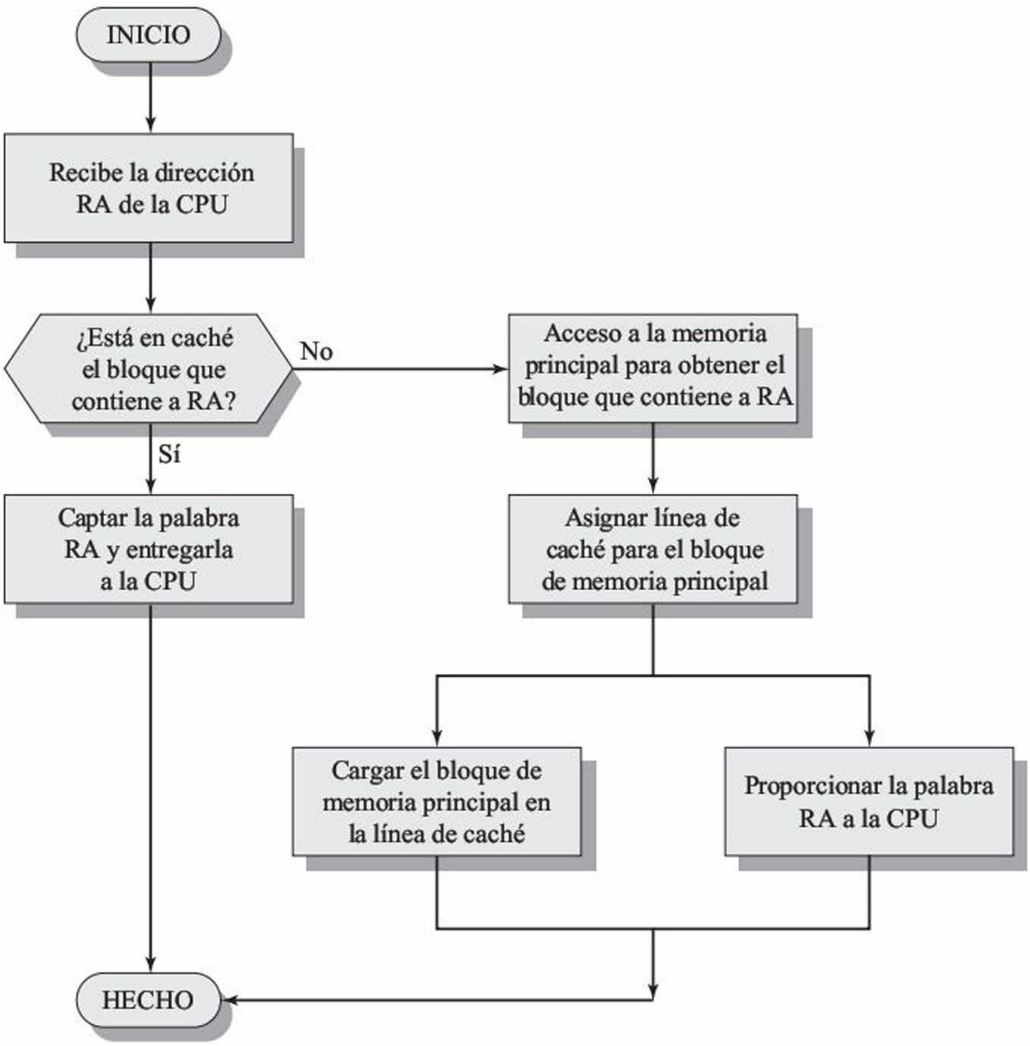

Figura 4.5. Operación de lectura de caché.

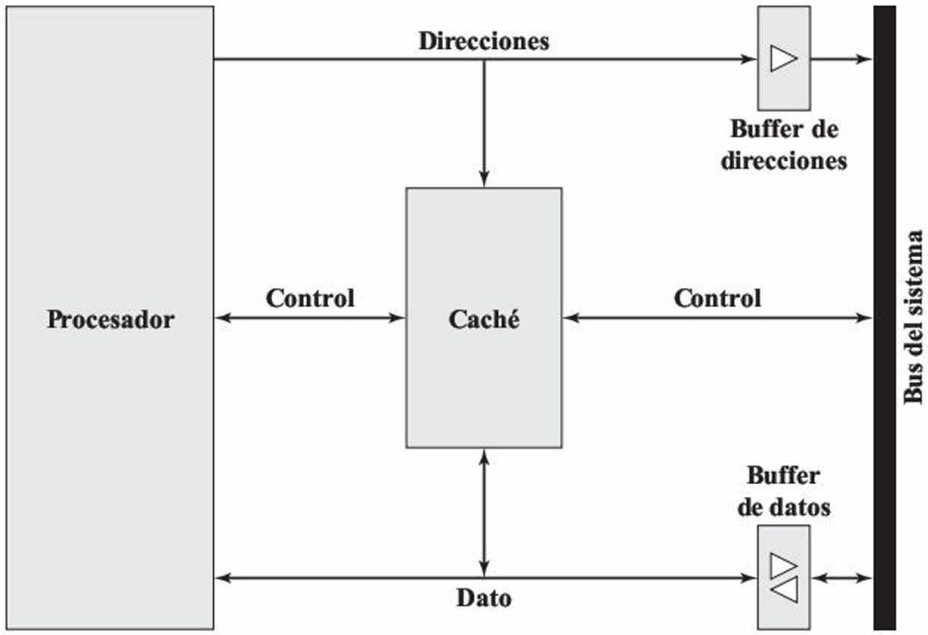

Figura 4.6. Organización típica de caché.

principal para todas las líneas de datos, direcciones y control. En este caso, frente a un fallo de caché, la palabra deseada es primero leída por la caché y después transferida desde esta al procesador.

El Apéndice 4A contiene un análisis de los parámetros de prestaciones relativos al uso de la caché.

4.3. ELEMENTOS DE DISEÑO DE LA CACHÉ

En esta sección se revisan los parámetros de diseño de la caché y se muestran algunos resultados típicos. A veces nos referimos al uso de cachés en el contexto de la computación de altas prestaciones (HPC, *High Performancea Computing*). La HPC trata los supercomputadores y su programación, especialmente para aplicaciones científicas que implican grandes cantidades de datos, cálculos con vectores y matrices, y el uso de algoritmos paralelos. El diseño de cachés para HPC difiere bastante del diseño para otras plataformas hardware y aplicaciones. Realmente, diversos investigadores han concluido que las aplicaciones de HPC presentan unas prestaciones pobres en arquitecturas de computadores que emplean cachés [BAIL93]. Desde entonces, otros investigadores han mostrado que una jerarquía de cachés puede ser útil para mejorar las prestaciones si el software de aplicación permite una explotación adecuada de la caché [WANG99, PRES01][3].

Aunque hay muy diversas implementaciones de caché, existen unos cuantos criterios básicos de diseño que sirven para clasificar y diferenciar entre arquitecturas de caché. La Tabla 4.2 lista algunos elementos clave.

[3] *Véase* [DOWD98] para un tratamiento más general de la HPC.