GOOD83 GOODMAN, J.: «Using Cache Memory to Reduce Processor-Memory Bandwidth». *Proceedings, 10th Annual International Symposium on Computer Architecture,* 1983. Reprinted in [HILL00].

HAND98 HANDY, J.: *The Cache Memory Book.* San Diego. Academic Press, 1993.

HIGB90 HIGBIE, L.: «Quick and Easy Cache Performance Analysis». *Computer Architecture News,* junio 1990.

HINT01 HINTON, G., *et al.:* «The Microarchitecture of the Pentium 4 Processor». *Intel Technology Journal,* Q1 2001. **http://developer.intel.com/technology/itj/**

SMIT82 SMITH, A.: «Cache Memories». *ACM Computing Surveys,* septiembre, 1992.

WILK65 WILKES, M.: «Slave memories and dynamic storage allocation». *IEEE Transactions on Electronic Computers,* abril, 1965. Reimpreso en [HILL00].

4.6. PALABRAS CLAVE, PREGUNTAS DE REPASO Y PROBLEMAS

PALABRAS CLAVE

acceso aleatorio	caché separada o partida	fallo de caché
acceso directo	caché unificada	jerarquía de memoria
acceso secuencial	computación de altas prestaciones (HPC)	línea de caché
acierto de caché	conjunto de caché	localidad
algoritmo de sustitución	correspondencia asociativa	localidad espacial
caché de datos	correspondencia asociativa por conjuntos	localidad temporal
caché de instrucciones	correspondencia directa	memoria caché
caché l3	escritura inmediata	postescritura
caché L1	escritura única	tasa de aciertos
caché L2	etiqueta	tiempo de acceso
caché multinivel		

PREGUNTAS DE REPASO

4.1. ¿Qué diferencias hay entre acceso secuencial, acceso directo y acceso aleatorio?

4.2. ¿Cuál es la relación general entre tiempo de acceso, coste y capacidad de memoria?

4.3. ¿Cómo se relaciona el principio de localidad con el uso de múltiples niveles de memoria?

4.4. ¿Qué diferencias existen entre las correspondencias directa, asociativa y asociativa por conjuntos?

4.5. Para una caché con correspondencia directa, una dirección de memoria principal es vista como tres campos. Enumere y defina estos campos.

4.6. Para una caché con correspondencia asociativa, una dirección de memoria principal es vista como dos campos. Enumere y defina estos campos.

4.7. Para una caché con correspondencia asociativa por conjuntos, una dirección de memoria principal es vista como tres campos. Enumere y defina estos campos.

4.8. ¿Qué diferencia hay entre localidad espacial y localidad temporal?

4.9. En general, ¿cuáles son las estrategias para explotar la localidad espacial y la localidad temporal?

PROBLEMAS

4.1. Una caché asociativa por conjuntos consta de 64 líneas divididas en conjuntos de 4 líneas. La memoria principal contiene 4K bloques de 128 palabras cada uno. Muestre el formato de direcciones de memoria principal.

4.2. Una caché asociativa por conjuntos de dos vías tiene líneas de 16 bytes y una capacidad total de 8 KB. La memoria principal, de 64 MB, es direccionable por bytes. Muestre el formato de las direcciones de memoria principal.

4.3. Para las direcciones hexadecimales de memoria principal: 111111, 666666, BBBBBB; muestre en formato hexadecimal la siguiente información:

a) Los valores de etiqueta, línea y palabra para una caché con correspondencia directa, utilizando el formato de la Figura 4.8.

b) Los valores de etiqueta y de palabra para una caché asociativa, utilizando el formato de la Figura 4.10.

c) Los valores de etiqueta, conjunto y palabra para una caché asociativa por conjuntos de dos vías, utilizando el formato de la Figura 4.12.

4.4. Indique los siguiente valores:

a) Para la caché directa del ejemplo la Figura 4.8: la longitud de la dirección, el número de unidades direccionables, el tamaño de bloque, el número de bloques en memoria principal, el número de líneas en caché y el tamaño de la etiqueta.

b) Para la caché asociativa del ejemplo la Figura 4.10: la longitud de la dirección, el número de unidades direccionables, el tamaño de bloque, el número de bloques en memoria principal, el número de líneas en caché y el tamaño de la etiqueta.

c) Para la caché asociativa por conjuntos del ejemplo la Figura 4.12: la longitud de la dirección, el número de unidades direccionables, el tamaño de bloque, el número de bloques en memoria principal, el número de líneas en un conjunto, el número de líneas en caché y el tamaño de la etiqueta.

4.5. Considere un microprocesador de 32 bits que tiene una caché *on-chip* de 16 KBytes asociativa por conjuntos de cuatro vías. Suponga que la caché tiene un tamaño de línea de cuatro palabras de 32 bits. Dibuje un diagrama de bloques de esta caché mostrando su organización y cómo se utilizan los diferentes campos de dirección para determinar un acierto/fallo de caché. ¿Dónde se asigna, dentro de la caché, la palabra de la posición de memoria ABCDE8F8?

4.6. Dadas las siguientes especificaciones para una memoria caché externa: asociativa por conjuntos de cuatro vías; tamaño de línea de dos palabras de 16 bits; capaz de albergar un total de 4K palabras de 32 bits de la memoria principal; utilizada con un procesador de 16 bits que emite direcciones de 24 bits. Diseñe la estructura de caché con toda la información pertinente y muestre cómo interpreta las direcciones del procesador.

4.7. El Intel 80486 tiene una caché unificada *on-chip*. Esta caché es de 8 KB y tiene una organización asociativa por conjuntos de cuatro vías y una longitud de bloque de cuatro palabras de 32 bits. La caché está estructurada en 128 conjuntos. Hay un único «bit de línea válida» y tres bits, B0, B1, y B2 (los bits de LRU), por línea. En un fallo de caché, el 80486 lee una línea de 16 bytes de memoria principal en una ráfaga de lectura de memoria a través del bus. Dibuje un diagrama simplificado de la caché, y muestre cómo son interpretados los diferentes campos de la dirección.

4.8. Considere una máquina con una memoria principal de 2^{16} bytes, direccionable por bytes, y un tamaño de bloque de 8 bytes. Suponga que con esta máquina se utiliza una caché de 32 líneas y correspondencia directa.

a) ¿Cómo se divide la dirección de memoria de 16 bits entre etiqueta, número de línea y número de byte?

b) ¿En qué líneas se almacenarían los bytes que se encuentran en las siguientes direcciones?
0001 0001 0001 1011
1100 0011 0011 0100
1101 0000 0001 1101
1010 1010 1010 1010

c) Suponga que se almacena en la caché el byte de dirección 0001 1010 0001 1010. ¿Cuáles son las direcciones de los bytes que se almacenan junto con él?

d) ¿Cuántos bytes de memoria pueden almacenarse en total en la caché?

e) ¿Por qué se almacenan también las etiquetas en la caché?

4.9. Para su caché *on-chip*, el Intel 486 utiliza un algoritmo de sustitución denominado pseudo LRU. Asociados con cada uno de los 128 conjuntos de cuatro líneas (etiquetadas L0, L1, L2, L3) hay tres bits, B0, B1, y B2. El algoritmo de sustitución opera así: cuando se debe sustituir una línea, la caché determinará primero si el uso más reciente fue de L0 y L1 o de L2 y L3. Entonces la caché determinará cuál de la pareja de bloques fue utilizado menos recientemente y lo marcará para sustituirlo. La figura 4.15 muestra la lógica asociada.

a) Especifique cómo se ponen los bits B0, B1 y B2, y cómo se utilizan estos en el algoritmo de sustitución de la Figura 4.15.

b) Muestre cómo el algoritmo del 80486 aproxima a un algoritmo LRU verdadero. *Sugerencia:* considere el caso en el que el orden de uso más reciente es L0, L2, L3, L1.

c) Demuestre que un algoritmo LRU verdadero requeriría seis bits por conjunto.

4.10. Una caché asociativa por conjuntos tiene un tamaño de bloque de cuatro palabras de 16 bits y un tamaño de conjunto de 2. La caché puede albergar un total de 4096 palabras. El tamaño de memoria principal que es transferible a caché es de 64K × 32 bits. Diseñe la estructura de caché y muestre cómo son interpretadas las direcciones del procesador.

4.11. Considere un sistema de memoria que emplea direcciones de 32 bits para direccionar a nivel de bytes, más una caché que usa un tamaño de línea de 64 bytes.

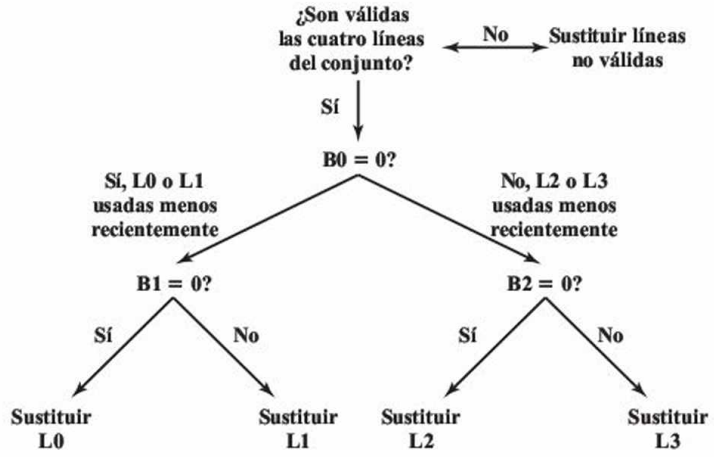

Figura 4.15. Estrategia de sustitución de la caché *on-chip* del 80486.

a) Suponga una caché con correspondencia directa, con un campo de etiqueta en la dirección de veinte bits. Muestre el formato de direcciones y determine los siguientes parámetros: número de unidades direccionables, número de bloques en memoria principal, número de líneas en caché y tamaño de la etiqueta.

b) Suponga una caché asociativa. Muestre el formato de las direcciones y determine los siguientes parámetros: número de unidades direccionables, número de bloques en memoria principal, número de líneas en caché y tamaño de la etiqueta.

c) Suponga una caché asociativa por conjuntos de 4 vías con un campo de etiqueta en la dirección de 9 bits. Muestre el formato de las direcciones y determine los siguientes parámetros: número de unidades direccionables, número de bloques en memoria principal, número de líneas en un conjunto, número de conjuntos en caché, número de lineas en caché y tamaño de la etiqueta.

4.12. Considere un computador con las siguientes características: un total de 1 MB de memoria principal; el tamaño de palabra es de un byte; el tamaño de bloque es de 16 bytes; y un tamaño de caché de 64 KB.

a) Para las direcciones de memoria principal: F0010, 01234, y CABBE, indique las correspondientes etiquetas, dirección de línea de caché y desplazamientos de palabras para una caché con correspondencia directa.

b) Indique dos direcciones cualesquiera de memoria principal con etiquetas diferentes que se correspondan con la misma línea para una caché con correspondencia directa.

c) Para las direcciones de memoria principal: F0010 y CABBE, indique los valores de etiqueta y de desplazamiento para una caché totalmente asociativa.

d) Para las direcciones de memoria principal: F0010 y CABBE, indique los valores de etiqueta, de conjunto de caché y de desplazamiento para una caché asociativa por conjuntos de dos vías.

4.13. Describa una técnica sencilla para implementar un algoritmo de sustitución LRU en una caché asociativa por conjuntos de cuatro vías.

4.14. Considere de nuevo el Ejemplo 4.3. ¿Cómo cambia el resultado si la memoria principal usa una capacidad de transferencia en bloques que tiene un tiempo de acceso de 30 ns a la primera de las palabras, y de 5 ns para cada una de las siguientes?

4.15. Considere el siguiente código:

```
for (i = 0; i < 20; i++)
for (j = 0; j < 10; j++)
a[i] = a[i] * j
```

a) Indique un ejemplo de localidad espacial en el código.

b) Indique un ejemplo de localidad temporal en el código.

4.16 Generalice las ecuaciones (4.1) y (4.2), del Apéndice 4A, a jerarquías de memoria de N niveles.

4.17. Un computador tiene una memoria principal de 32K palabras de 16 bits. Tiene también una caché de 4K palabras dividida en conjuntos de cuatro líneas con 64 palabras por línea. Suponga que la caché está inicialmente vacía. El procesador capta palabras de las posiciones 0, 1, 2..., 4351, en ese orden. Entonces repite esta secuencia de captación nueve veces más. La caché es diez veces más rápida que la memoria principal. Estime la mejora resultante por el uso de la caché. Suponga una política LRU para la sustitución de bloques.

4.18. Considere una caché de cuatro líneas con 16 bytes cada una. La memoria principal está dividida en bloques de 16 bytes. Es decir el bloque 0 tiene bytes con direcciones 0 a 15, y así sucesivamente. Considere ahora un programa que accede a memoria con la siguiente secuencia de direcciones:

Una vez: de 63 a 70.

Diez veces en un bucle: de 15 a la 32; y 80 a 95.

a) Suponga que la caché es de correspondencia directa. Los bloques de memoria 0, 4, etc., se asignan en la línea 1; los bloques 1, 5, etc., en la línea 2, y así sucesivamente. Calcule la tasa de aciertos.

b) Suponga ahora que la caché tiene una organización asociativa por conjuntos de 2 vías, on dos conjuntos de dos líneas cada uno. Los bloques con numeración par se asignan al conjunto 0 y los impares al conjunto 1. Calcule la tasa de aciertos para la caché asociatva por conjuntos de dos vías usando el esquema de sustitución LRU.

4.19. Considere un sistema de memoria con los siguientes parámetros:

$$T_c = 100 \text{ ns} \qquad C_c = 10^{-4} \text{ dólares/bit}$$
$$T_m = 1.200 \text{ ns} \qquad C_m = 10^{-5} \text{ dólares/bit}$$

a) ¿Cuál es el coste de una memoria principal de 1 MB?

b) ¿Cuál es el coste de una memoria principal de 1 MB utilizando la tecnología de la caché?

c) Si el tiempo de acceso efectivo es un 10 por ciento mayor que el tiempo de acceso de la caché, ¿cuál es la tasa de aciertos H?

4.20. a) Considere una caché L1 con un tiempo de acceso de 1 ns y una tasa de aciertos $H = 0{,}95$. Suponga que queremos cambiar el diseño de la caché (el tamaño, su organización) de manera que incrementemos H hasta 0,97, pero aumentando el tiempo de acceso a 1,5 ns. ¿Qué condiciones deben cumplirse para que este cambio suponga una mejora en las prestaciones?

b) Explique por qué el resultado tiene sentido intuitivamente.

4.21. Considere una caché de un solo nivel, con un tiempo de acceso de 2,5 ns, un tamaño de línea de 64 bytes y una tasa de aciertos $H = 0{,}95$. La memoria principal usa la capacidad de transferencia en bloques, con un tiempo de acceso de 50 ns para la primera palabra (4 bytes), y de 5 ns para cada una de las siguientes.

a) ¿Qué valor tiene el tiempo de acceso cuando hay un fallo de caché? Suponga que la caché espera hasta que la línea ha sido captada de memoria principal, para entonces ejecutar un acierto de caché.

b) Suponga que al incrementar el tamaño de línea a 128 bytes se incrementa H hasta 0,97. ¿Reduce esto el tiempo medio de acceso a memoria?

4.22. Un computador dispone de una caché, memoria principal, y un disco utilizado para memoria virtual. Cuando se referencia una palabra que está en la caché se requieren 20 ns para acceder a ella. Si está en memoria principal pero no en la caché se necesitan 60 ns para cargarla en la caché, y entonces se inicia de nuevo la referencia. Si la palabra no está en memoria principal se necesitan 12 ms para captarla de disco, seguidos de 60 ns para copiarla en la caché, comenzando entonces de nuevo la referencia. La tasa de aciertos de caché es de 0,9 y la de memoria principal de 0,6. ¿Cuál es, en nanosegundos, el tiempo medio necesario para acceder a una palabra referenciada en este sistema?

4.23. Considere una caché con un tamaño de línea de 64 bytes. Suponga que, en media, un 30 por ciento de las líneas de caché son modificadas. Una palabra consta de 8 bytes.

a) Suponga una tasa de fallos del 3 por ciento (tasa de aciertos de 0,97). Calcule la cantidad de tráfico de memoria principal en términos de bytes por instrucción, para políticas de escritura inmediata y de postescritura. Las lecturas de memoria principal a caché se realizan de línea en línea. Sin embargo, para la postescritura puede escribirse una sola palabra de caché a memoria principal.

b) Repita el apartado **a** para una tasa del 5 por ciento.

c) Repita el apartado **a** para una tasa del 7 por ciento.

d) ¿Qué conclusión puede extraerse de los resultados?

4.24. En el microprocesador Motorola 68020, un acceso a caché ocupa dos ciclos de reloj. El acceso a datos desde memoria principal, a través del bus, hasta el procesador, ocupa tres ciclos de reloj incluso cuando no se inserten estados de espera; los datos se entregan al procesador a la vez que se entregan a la caché.

a) Calcule la duración efectiva de un ciclo de memoria para una tasa de aciertos de 0,9 y una frecuencia de reloj de 16,67 MHz.

b) Repita los cálculos suponiendo que se insertan dos estados de espera de un ciclo por cada ciclo de memoria. ¿Qué conclusión puede extraerse de estos resultados?

4.25. Un procesador tiene un tiempo de ciclo de memoria de 300 ns y una velocidad de procesamiento de instrucciones de 1 MIPS. De media, cada instrucción necesita un ciclo de memoria del bus para captar la instrucción y otro para el operando involucrado.

a) Calcule la utilización del bus por parte del procesador.

b) Suponga que el procesador dispone de una caché de instrucciones con una tasa de aciertos asociada de 0,5. Determine el efecto que tiene sobre la utilización del bus.

4.26. Las prestaciones de un sistema de caché de solo un nivel, para una operación de lectura, puede caracterizarse mediante la ecuación:

$$T_a = T_c + (1 - H)\, T_m$$

Donde T_a es el tiempo de acceso medio, T_c es el tiempo de acceso a caché, T_m es el tiempo de acceso a memoria (de memoria a registro del procesador), y H es la tasa de aciertos. Para simplificar suponemos que la palabra en cuestión se carga en la caché en paralelo con su carga en el registro del procesador. Tiene la misma forma que la Ecuación (4.1).

a) Defina T_b = tiempo de transferencia de una línea entre caché y memoria principal, y W = fracción de referencias para escritura. Revise la ecuación anterior para que tenga en cuenta tanto las escrituras como las lecturas, usando una política de escritura inmediata.

b) Defina W_b como la probabilidad de una línea en caché haya sido modificada. Obtenga una ecuación para T_a, con una política de postescritura.

4.27. Para un sistema con dos niveles de caché, defina T_{c1} = tiempo de acceso a la caché del primer nivel; T_{c2} = tiempo de acceso a la caché del segundo nivel; T_m = tiempo de acceso a memoria; H_1 = tasa de aciertos de la caché del primer nivel; H_2 = tasa aciertos combinada del primer y el segundo nivel. Obtenga la ecuación de T_a para una operación de lectura.

4.28. Suponga el siguiente comportamiento frente a un fallo de caché: un ciclo de reloj para enviar una dirección a la memoria principal y cuatro ciclos de reloj para acceder a una palabra de 32 bits de la memoria principal y transferirla al procesador y a la caché.

a) Si el tamaño de línea de caché es de una palabra, ¿cuál es la penalización por fallo (es decir, el tiempo adicional necesario para una lectura cuando se produce un fallo de lectura)?

b) ¿Cuál es la penalización por fallo si el tamaño de línea de caché es de cuatro palabras, y se ejecuta una transferencia múltiple, no en ráfaga?

c) ¿Cuál es la penalización por fallo si el tamaño de línea de caché es de cuatro palabras y se ejecuta una transferencia, con un pulso de reloj para transferir cada palabra?

4.29. Para el diseño de caché del problema anterior, suponga que el incremento del tamaño de línea de una a cuatro palabras produce una disminución de la tasa de fallos de lectura del 3,2 por ciento al 1,1 por ciento. Para ambos casos, transferencia en ráfagas o no, ¿cuál es la penalización por fallo media, promediada sobre todas las lecturas, para los dos tamaños de línea indicados?

APÉNDICE 4A. PRESTACIONES DE LAS MEMORIAS DE DOS NIVELES

En este capítulo se ha hecho referencia a la caché que actúa como buffer entre la memoria principal y el procesador, creando una memoria interna de dos niveles. Esta arquitectura de dos niveles proporciona mejores prestaciones que una memoria comparable de un solo nivel, explotando una propiedad conocida como localidad, que analizamos más adelante en este apéndice.

El mecanismo de caché de la memoria principal es parte de la arquitectura del computador implementada en hardware y normalmente invisible para el sistema operativo. Además, hay otros dos ejemplos de memorias de dos niveles que también aprovechan la localidad y que se implementan, al menos parcialmente, en el sistema operativo: la memoria virtual y la caché de disco (Tabla 4.7). La