La Enseñanza sobre Memorias Caché en la Titulación de Ingeniería Informática de la Universidad de Extremadura

Miguel A. Vega Rodríguez, Juan A. Gómez Pulido, Juan M. Sánchez Pérez
Departamento de Informática. Universidad de Extremadura
Escuela Politécnica. Campus Universitario, s/n. 10071 Cáceres. Spain
E-mail: mavega@unex.es, jangomez@unex.es, sanperez@unex.es. Fax: +34-927-257202

Resumen

Con este trabajo se intenta responder a la pregunta de cómo se enseña memoria caché en el plan de estudios de la titulación de Ingeniería Informática de la Universidad de Extremadura. La docencia de memorias caché se distribuye entre dos asignaturas obligatorias de 2º y 4º curso: Estructura de Computadores y Arquitectura de Computadores. En este trabajo se detallan los objetivos generales y específicos de cada asignatura, las horas dedicadas al estudio de memorias caché, la bibliografía recomendada, etc.

1. Introducción

En el plan de estudios de la titulación de Ingeniería Informática de la Universidad de Extremadura (UEX), destacan dos asignaturas obligatorias, anuales y de caracter fundamental dentro del área de *Arquitectura y Tecnología de Computadores*:

- *EC*: Estructura de Computadores (2º curso).
- *AC*: Arquitectura de Computadores (4º curso).

El estudio de las memorias caché dentro de esta titulación se distribuye entre estas dos asignaturas. En EC se estudia la memoria caché por primera vez, detallando desde sus principios básicos hasta las características de los sistemas caché multinivel. En la asignatura AC, partiendo de la base de los conocimientos adquiridos en EC, se amplía el estudio de la memoria caché al caso de los sistemas multiprocesador.

2. Estructura de Computadores

La asignatura de Estructura de Computadores es una asignatura obligatoria anual de 12 créditos (6 teóricos y 6 prácticos), lo que contabiliza un total de 120 horas, de las cuales 60 son de teoría (2 horas teóricas por semana). Se organiza con 2 grupos de teoría (de unos 90 alumnos cada uno) y 6 de prácticas (de unos 30 alumnos cada uno). Tiene como descriptores los siguientes items:

- Instrucciones: formatos y direccionamiento.
- Sistema de memoria.
- Unidad de control.
- Sistema E/S.
- Buses.
- Aumento de prestaciones.

Con la guía de los descriptores, se ha elaborado el contenido teórico de EC mediante 6 capítulos:

- 1: Introducción.
- 2: Arquitectura del conjunto de instrucciones.
- 3: Aritmética de procesadores.
- 4: Sistema procesador.
- 5: Sistema de memoria.
- 6: Sistemas de E/S y buses.

El tema de las memorias caché corresponde, junto con los temas de memoria principal y memoria virtual, con el descriptor "sistema de memoria", que se materializa en el capítulo 5º (sistema de memoria). El capítulo completo gira en torno a un sistema de jerarquía de memoria, profundizando en la problemática de la memoria desde la perspectiva de la organización funcional, con vistas a conseguir un mayor rendimiento.

Para el caso concreto del tema de las memorias caché, dentro del capítulo 5º, se sigue el siguiente programa:

- Definiciones generales. Principio de localidad. Localidad temporal. Localidad espacial. Niveles. Bloque. Acierto. Fallo. Frecuencia de aciertos. Frecuencia de fallos. Tiempo de acierto. Penalización de fallo. Tiempo de acceso. Tiempo de transferencia. Tiempo de fallo. Dirección de memoria. Dirección de la estructura de bloque. Dirección del desplazamiento de bloque. Tiempo medio de acceso. Tiempo de ciclo.
- Concepto de caché. Fundamento de la existencia de la caché.
- Aplicación a la caché de los conceptos de jerarquía de memoria. Nivel superior e inferior. Acceso o referencia a memoria.
- Organización de la caché. Bloque. Etiqueta. Bit de validez. Dirección de memoria. Valores típicos. Ejemplo.
- Correspondencia. Correspondencia directa. Totalmente asociativa. Asociativa por conjuntos de N vías.
- Reemplazamiento. Aleatorio. Menos recientemente usado (LRU). Primero en entrar, primero en salir (FIFO). Menos frecuentemente usado (LFU).
- Operación de la caché. Inicialización. Lectura. Escritura. Escritura directa. Postescritura.
- Problemática de los fallos. Tipos de fallos. Forzosos. Capacidad. Conflicto. Diseño para evitar estos fallos.
- Cachés unificadas y separadas. Unificadas o mixtas. Separadas. Ventajas. Rendimiento.
- Rendimiento de la caché. Evaluación del rendimiento. Recuento de instrucciones. Frecuencia de fallos. Tiempo medio de acceso a memoria. Tiempo de ejecución: Definición. Número total de instrucciones. Periodo o ciclo de reloj. Ciclos de reloj de ejecución de CPU por instrucción. Ciclos de reloj de detención de CPU por instrucción: Fórmula. Conclusiones.
- Mejora del rendimiento. Tamaño de bloque. Respecto a la frecuencia de fallos. Respecto a la penalización de fallos. Respecto al tiempo medio de acceso (TMA). Consideración final. Cachés multinivel: Definiciones, esquema. TMA para sistema caché multinivel. Diseño de la caché de 2º nivel. Mejora del rendimiento. Inclusión multinivel. Parámetros típicos en cachés de 2º nivel. Mejora de la localidad de los programas: Organización, tamaño de la caché. Otras formas de mejorar el rendimiento.
- Tipos de cachés en procesadores actuales.

Para conseguir una estructura teórica uniforme, dentro del capítulo 5º, se comienza por el tema que trata de la jerarquía de memoria. Se introducen los fundamentos y conceptos que se emplean en todo el capítulo (niveles, bloque, aciertos, tiempos, direcciones de memoria, etc). Después se estudia en profundidad la memoria caché. Este tema comienza con una introducción que define para el caso de la caché: conceptos, organización, dirección de memoria, valores típicos de los parámetros, etc. Una vez conseguido esto, se abordan los algoritmos de correspondencia y las estrategias de reemplazamiento. Con toda esta información, se pueden ya explicar las distintas operaciones de caché (inicialización, lectura, escritura), con toda la problemática asociada. De todo ello se comprenderá la importancia en reducir los fallos de caché, para lo cual presentamos un apartado que aborda toda la problemática asociada a los fallos (tipos, diseño para evitarlos, etc.). También se describen las cachés unificadas y separadas para evaluar su impacto diferenciado en los tipos de fallos. El rendimiento y la mejora de éste se dejan para sendos apartados en los que se estudia todo lo asociado con el aumento de prestaciones, haciendo especial referencia en los parámetros que puede cambiar el diseñador con ese propósito.

El resto del capítulo apuntala la memoria caché dentro de la jerarquía del sistema de memoria. Así, se estudia la memoria principal y la virtual. En total, el capítulo dedica 12 horas (6 semanas) al estudio de sus temas. La memoria caché consume 8 horas (4 semanas), convirtiéndose así en el centro de gravedad del capítulo. Esto es así, entre otras cosas, porque la memoria principal no requiere de grandes extensiones teóricas, y porque la memoria virtual es estudiada también por la asignatura troncal de Sistemas Operativos.

Como metodología docente, indicamos que el capítulo se explica mediante transparencias, que sintetizan la información recopilada de varios textos, principalmente de [5]. Este texto junto con [3], ambos de los mismos autores, son la

bibliografía básica recomendada. Los dos poseen una calidad y prestigio ampliamente demostrados, incluyendo explicaciones teóricas muy completas y adecuadas, acompañadas por un gran número de gráficas que permiten aclarar los aspectos más complejos (p.e. la interrelación entre los distintos parámetros de una caché). Además, es posible encontrar ambos textos traducidos al español.

Finalmente, indicar que la teoría sobre cachés se ve complementada por varias prácticas que se realizan, dentro de la asignatura, durante la primera mitad del segundo cuatrimestre. Para la realización de estas prácticas se utiliza el simulador SISMEC [2], construido por los profesores de la asignatura especialmente para este fin.

3. Arquitectura de Computadores

La asignatura Arquitectura de Computadores es una asignatura troncal de 9 créditos (6 teóricos y 3 prácticos), que corresponden a 90 horas lectivas (60 teóricas y 30 prácticas). Se organiza con 1 grupo de teoría (de unos 120 alumnos) y 4 de prácticas (de unos 30 alumnos cada uno). Para la elaboración de los contenidos de la asignatura se han tenido en cuenta los conocimientos adquiridos por los alumnos en las asignaturas de primer ciclo: Estructuras de Computadores, Sistemas Operativos y Diseño Automático de Sistemas, así como la relación de esta asignatura con las restantes asignaturas de segundo ciclo.

Los objetivos fundamentales que se pretenden conseguir con esta asignatura son:

- Introducir el concepto de Arquitectura de Computadores.
- Repasar los conceptos sobre jerarquía de memoria, y en particular sobre cachés.
- Introducir el concepto de paralelismo a distintos niveles.
- Estudiar las técnicas de segmentación.
- Introducir el concepto de procesador vectorial.
- Estudiar las características de los sistemas multiprocesadores, haciendo especial énfasis en su organización de memoria.
- Introducir el concepto de multicomputador, y los mecanismos de paso de mensajes.

El programa de la asignatura está dividido en una serie de módulos con los que se pretenden alcanzar los objetivos antes mencionados. La relación de módulos es la siguiente:

I. Fundamentos de la Arquitectura de Computadores.
II. Gestión de la jerarquía de memoria.
III. Introducción al procesamiento paralelo.
IV. Segmentación.
V. Procesadores vectoriales segmentados.
VI. Multiprocesadores.
VII. Multicomputadores.

En este trabajo nos centraremos únicamente en los módulos II y VI, por ser en ellos donde se realiza el estudio sobre memorias caché, tanto en sistemas uniprocesador (módulo II) como multiprocesador (módulo VI). Cada uno de estos módulos se descompone en un conjunto de temas. Indicaremos para cada tema la duración, objetivos, contenidos, así como un breve resumen; además de la bibliografía recomendada.

3.1. Repasando la jerarquía de memoria

El módulo II del programa de la asignatura se dedica al repaso de la jerarquía de memoria, y en particular a refrescar los conceptos fundamentales sobre memorias caché. Todos estos aspectos fueron estudiados por los alumnos en la asignatura EC, impartida en segundo curso de carrera. Hemos optado por su repaso para que los alumnos estén totalmente preparados a la hora de enfrentarse al estudio de memorias caché dentro de multiprocesadores, que se realiza en el módulo VI.

El módulo II se encuentra dividido en dos temas. El primero de ellos lleva por título "*Organización Jerárquica de la Memoria*", y con él se busca:

- Repasar los conceptos y objetivos de la jerarquía de memoria.
- Explicar las técnicas básicas para crear un gran espacio de direcciones virtuales.

Para conseguir estos objetivos el tema se ha dividido en los siguientes apartados:

1. Conceptos básicos.

2. Optimización de la jerarquía de memoria.
3. Memoria virtual.
4. Memoria paginada.
5. Memoria segmentada.

Se comienza la lección explicando los principios básicos de la jerarquía de memoria, presentando también el problema de la optimización de los tiempos de acceso de la jerarquía, sujetos a un coste del sistema de memoria. A continuación, se introduce el concepto de memoria virtual presentando la problemática de la organización del espacio de direcciones de memoria. Se estudian posteriormente dos métodos de implementación del espacio de direcciones: paginación y segmentación. Debido a que todos estos aspectos ya fueron estudiados por los alumnos, se dedica aproximadamente un tiempo de una hora de teoría y una de problemas para ver esta lección de repaso.

El segundo y último tema del módulo II, denominado "*Memoria Caché*", tiene por objetivos:

- Repasar el esquema de funcionamiento de la memoria caché.
- Explicar las políticas de correspondencia.
- Estudiar políticas de actualización de memoria principal.
- Presentar diversas políticas de reemplazo de bloques.

Para conseguir estos objetivos el tema se encuentra estructurado en los siguientes apartados:

1. Características de las memorias caché.
2. Organización de las memorias caché.
3. Políticas de extracción y actualización de memoria principal.
4. Políticas de reemplazo de bloques.

En esta lección se repasan las características de las memorias cachés y las diferentes estrategias de administración de estas memorias. Se examinan cuatro organizaciones de memoria caché: directa, totalmente asociativa, asociativa por conjuntos y correspondencia por sectores. También se abordan las distintas políticas de reemplazo de bloques en la memoria caché (aleatoria, LRU, LFU, FIFO). Al igual que ocurría con el tema 1, al tratarse de una lección de repaso, se emplea aproximadamente un tiempo de una hora de teoría y una de problemas.

En conclusión, para el módulo II se dedican un total de dos horas de teoría y dos de problemas. Es posible utilizar un número tan reducido de horas debido a dos motivos. Por un lado, todos estos conceptos ya fueron estudiados por los alumnos con anterioridad, por lo que sólo se busca su actualización. Por otro lado, al alumno se le entrega una copia de las transparencias seguidas en clase. Es decir, el alumno dispone del material suficiente para no estar constantemente copiando las explicaciones del profesor, dando lugar a clases más ágiles. La bibliografía básica recomendada para este módulo es: [3] y [5]; por los mismos motivos que se aconsejan en la asignatura EC.

3.2. Memorias caché y multiprocesadores

El módulo VI del programa de la asignatura se dedica al estudio de las principales características de los sistemas multiprocesador, haciendo especial hincapié en su organización de memoria. El módulo se encuentra dividido en los siguientes bloques temáticos:

1. Introducción.
2. Redes de interconexión.
3. Sincronización de procesos.
4. Organización de memoria en multiprocesadores.
5. Aspectos de programación.

Nos centraremos aquí únicamente en el bloque temático 4, dedicado a la organización de memoria. Con este bloque se busca:

- Indicar las posibles jerarquías de memoria en multiprocesadores.
- Presentar en detalle el problema de la coherencia caché.
- Introducir los esquemas de coherencia caché basados en software.
- Estudiar los esquemas de coherencia caché basados en hardware, diferenciando entre arquitecturas de red con caché coherente y protocolos de coherencia caché.

- Distinguir entre protocolos de coherencia caché basados en directorio y de espionaje (*snoopy*).
- Exponer varios ejemplos de protocolos de espionaje.
- Considerar en detalle (nivel de transición de estados) algunos de estos ejemplos.

Para conseguir estos objetivos el tema se ha dividido en los siguientes apartados:

1. Jerarquías de memoria en multiprocesadores.
2. La caché y el problema de la coherencia.
3. Esquemas de coherencia caché software y hardware.
4. Protocolos hardware de coherencia caché.
5. Los protocolos de espionaje. Tres casos de estudio.

La lección comienza con una introducción sobre las posibles jerarquías de memoria en multiprocesadores. Se muestra que de las categorías existentes la más interesante es la de los multiprocesadores con memoria compartida por bus. De hecho, es el modelo dominante de máquina paralela que se vende en la actualidad [1]. Éste es el motivo de que nos centremos en los esquemas de coherencia caché para estos sistemas. Posteriormente se describe el problema de la coherencia de caché, realizando un estudio detallado. Se muestra la existencia de esquemas de coherencia caché basados en software y en hardware. Aunque relatamos varios esquemas de coherencia caché basados en software nos centramos en los basados en hardware. Dentro de los esquemas de coherencia basados en hardware diferenciamos entre arquitecturas de red con caché coherente y protocolos de coherencia caché. Las arquitecturas de red quedan fuera de nuestros objetivos, por lo que estudiamos más en detalle los protocolos de coherencia caché. Hay dos clases de protocolos: los basados en directorio y los de espionaje (*snoopy*). Tras dar las bases de ambos, razonamos porqué vamos a estudiar más detalladamente los protocolos de espionaje. Llegamos a exponer un total de siete protocolos de espionaje distintos. En lugar de estudiar en profundidad todas las posibles elecciones, consideramos los tres protocolos de coherencia habituales (MSI, MESI o Illinois, y Dragon), que mostrarán las opciones de diseño. Para cada uno de estos protocolos describimos, a nivel de transición de estados, cómo trabajan.

Para este tema se dedica aproximadamente un tiempo de cuatro horas de teoría y una de problemas. Al igual que ocurría para el módulo II, el número de horas se ve reducido por el hecho de que al alumno se le entrega una copia de las transparencias seguidas en clase; Lo cual permite que el alumno se centre en comprender las explicaciones, y que el profesor disponga de más tiempo para aclarar, debatir, trabajar con el grupo, etc. El número de horas de problemas es tan reducido por realizarse, dentro de la asignatura y durante el segundo cuatrimestre, varias prácticas sobre sistemas de memoria caché en multiprocesadores. Siendo en estas prácticas donde se afianzan totalmente los conceptos aprendidos en teoría. Para la realización de las prácticas se utiliza el simulador SMPCache [2], construido por los profesores de la asignatura especialmente para este fin.

La bibliografía básica recomendada, sobre memorias caché en multiprocesadores, es: [1], [3], [4] y [6]. [1] es un texto de gran calidad donde se tratan en profundidad todos los aspectos sobre memorias cachés en multiprocesadores. De hecho, es el texto principal sobre el que se apoyan las clases teóricas. [3] es un texto de sobrado prestigio, de él destacamos tanto su estudio detallado de las memorias caché (acompañado de numerosas gráficas que muestran la influencia de diversos parámetros en la tasa de fallos), como la descripción que realiza de los protocolos de espionaje. [4] es un texto donde se explican en detalle los conceptos asociados con memorias caché, también se introduce el problema de la coherencia caché en multiprocesadores. Como principal desventaja resaltamos la antigüedad del texto. Finalmente, [6] tiene como principal punto fuerte la claridad de sus explicaciones, apoyadas en gran cantidad de gráficos. Además, en él no sólo tienen cabida los protocolos de espionaje sino que también se estudian en detalle los esquemas de directorio (al igual que ocurre en [1]). La mayoría de estos textos también se encuentran en versión española.

4. Conclusión

En este trabajo se ha indicado cómo se enseña memoria caché en la titulación de Ingeniería Informática de la Universidad de Extremadura. Se han repasado las asignaturas que incluyen docencia en memorias caché, y la importancia que ésta tiene dentro de los objetivos generales de cada asignatura. El número de horas dedicadas al estudio de memorias caché muestra que éstas son un aspecto fundamental dentro de todo sistema informático (uniprocesador o multiprocesador), principalmente porque son un componente crítico para el rendimiento de cualquier sistema.

Referencias

[1] Culler, D.E.; Singh, J.P.; Gupta, A. *Parallel Computer Architecture. A Hardware/Software Approach.* Morgan Kauffmann, 1999.

[2] GACDL. http://atc.unex.es/gacdl, 2001.

[3] Hennesy, J.L.; Patterson, D.A. *Computer Architecture - A Quantitative Approach*. Morgan Kauffmann, 1996.

[4] Hwang, K.; Briggs, F.A. *Computer Architecture and Parallel Processing*. McGraw-Hill, 1984.

[5] Patterson, D.A.; Hennesy, J.L. *Computer Organization and Design. The Hardware/Software Interface.* Morgan Kauffmann, 1994.

[6] Sima, D.; Fountain, T.; Kacsuk, P. *Advanced Computer Architectures. A Design Space Approach.* Addison-Wesley, 1998.