Universidad Autónoma del Estado de Hidalgo

Escuela Superior Huejutla

Area Académica: Sistemas Computacionales

Tema: Elementos de diseño de memoria caché

Profesor: Raúl Hernández Palacios

Periodo: 2011

Keywords: Memory, cache memory.

Tema: Elementos de diseño de memoria caché

Abstract: The use of the cache in a computer system increases the overall performance of it, so it is necessary to know the strategies for the design and implementation of this type of memory.

Keywords: Memory, cache memory.

Elementos de diseño de caché:

- **Tamaño de caché**
- **Función de correspondencia**
  - Directa
  - Asociativa
  - Asociativa por conjuntos
- **Algoritmo de sustitución**
  - Utilizado menos recientemente (LRU)
  - Primero en entrar primero en salir (FIFO)
  - Utilizado menos frecuentemente (LFU)
  - Aleatorio
- **Política de escritura**
  - Escritura inmediata
  - Postescritura
  - Escritura única
- **Tamaño de línea**
- **Número de cachés**
  - Uno o dos niveles
  - Unificada o partida

Tamaño de caché:

Se dice que entre más grande es el tamaño de caché, mayor es el número de puertas implicadas en direccionar la caché, por lo tanto cachés grandes son ligeramente más lentas que cachés pequeñas.

Tamaño de caché de procesadores:

Procesador	Año	Caché L1	Caché L2	Caché L3
IBM 3033	1978	64 KB	-	-
Intel 80486	1989	8 KB	-	-
Pentium	1993	8 KB	256 a 512 KB	-
PowerPC 64	1999	32 KB	256 KB / 1MB	2 MB
Pentium 4	2000	8 KB	256 KB	-
Itanium	2001	16 KB	96 KB	4 MB

Función de correspondencia:

- Debido a que existen menos líneas de caché que bloques de memoria principal, es necesario un algoritmo para hacer corresponder los bloques de memoria principal a las líneas de caché. De la misma forma es requerido un medio para determinar qué bloque de memoria principal ocupa en su momento una de línea dada de caché.

- Al elegir la función de correspondencia se determina la organización de caché.

- Técnicas: directa, asociativa y asociativa por conjuntos.

Algoritmos de sustitución:

- Para conseguir alta velocidad de comunicación es necesario que los algoritmos sean implementados en hardware.

- El algoritmo más efectivo actualmente es el "*utilizado menos recientemente" (LRU, least-recently used)*, es sustituido el bloque que se ha mantenido más tiempo en caché sin referenciarlo, el algoritmo es aplicado mediante bit para referenciar o no cada bloque

Algoritmos de sustitución:

- *Primero en entrar, primero en salir (FIFO).* Se sustituye el bloque que ha estado más tiempo en la caché, se aplica mediante la técnica round-robin, de forma cíclica.

- *Utilizado menos frecuentemente (LFU, least-frecuently used).* Se sustituye el bloque que ha sido menos referenciado, se implementa asociando un contador a cada línea de caché.

Política de escritura

Dos panoramas:

- Si un bloque antiguo en la caché no debe ser modificado, puede sobrescribirse con el nuevo bloque sin necesidad de actualizar el antiguo.

- Si se ha realizado por lo menos una operación de escritura sobre una palabra de la línea correspondiente de caché, entonces la memoria principal debe ser actualizada, rescribiendo la línea de caché en el bloque de memoria antes de transferir el nuevo bloque.

Política de escritura: Técnica – escritura inmediata.

Todas las operaciones de escritura son realizadas en la caché como en memoria principal, y asegura que el contenido de la memoria principal siempre será válido.

Para mantener la coherencia de caché cualquier otro módulo de procesador-caché puede monitorizar el tráfico, una desventaja es que se genera mucho tráfico con la memoria y puede originar un cuello de botella.

Política de escritura: Técnica – Postescritura.

Minimiza las escrituras en memoria. Todas las actualizaciones se realizan en la caché, cuando se tiene una actualización se activa un bit ACTUALIZAR asociado a la línea de caché, el bloque sólo es sustituido o escrito en memoria principal cuando se tiene activo el bit ACTUALIZAR.

Número de cachés:

Existen dos aspectos de diseño: número de niveles de caché y, usar cachés unidas a cachés separadas.

Cachés multinivel.- caché on-chip, reduce la actividad del bus externo del procesador, de esta forma se reduce el tiempo de ejecución y se incrementan las prestaciones globales del sistema. Cuando la instrucción se encuentra en la caché del procesador, se elimina el acceso al bus, de esta forma el bus queda libre para realizar otras transferencias.

Número de cachés:

Además es conveniente incluir un segundo nivel de caché, off-chip.

Caché de dos niveles.- on-chip y off-chip referente al procesador.

Porqué incluir caché L2:

Suponiendo que el procesador quiere acceder a una posición de memoria que no se encuentra en el L1, entonces el procesador accederá a la memoria principal mediante el bus, el bus es lento, se reducen prestaciones, por lo tanto, utilizar L2 ayuda a recuperar esa información faltante

Dos características de diseño de las cachés multinivel:

Primero.- Para caché L2.- algunos diseños no usan el bus del sistema para las transferencias entre el procesador y la caché L2, utilizan un camino de datos alterno, y, se reduce el tráfico en el bus de sistema.

Segundo.- al reducir las dimensiones de componentes del procesador, se puede incluir en el procesador el nivel L2 de caché, por consiguiente se mejora las prestaciones globales del sistema.

Dos características de diseño de las cachés multinivel:

Unión o separación de cachés.- al aparecer las cachés on-chip los diseños contenían almacenamiento de referencias tanto los datos como instrucciones. Actualmente se opta por la separación de cachés referente a: caché dedica a datos y caché dedicada a instrucciones.

Evolución de la caché en los procesadores Intel:

Problema	Solución	1er. Procesador en incluirla
Memoria externa más lenta que el bus del sistema	Añadir caché externa usando tecnología de memoria más rápida	386
El aumento de velocidad de los procesadores hace que el bus sea el cuello de botella para acceder a la caché	Trasladar la caché externa al mismo chip, funciona a la misma velocidad que el procesador	486
La caché es pequeña debido a la disponibilidad de superficie *on-chip* limitada	Añadir una caché L2 externa con tecnología más rápida que la usad en la memoria principal	486
Se produce contención cuando las unidades de precaptación de instrucciones y ejecución necesitan acceder simultáneamente a la caché. Precaptación se bloquea y ejecución accede a datos.	Crear cachés separadas de datos y de instrucciones.	Pentium
El incremento de velocidad del procesador hace que el bus externo sea un cuello de botella para el acceso a la caché L2	Crear un bus externo específico que funcione a la misma velocidad del bus principal.	Pentium Pro
	Llevar la caché L2 al chip del procesador	Pentium II
Algunas aplicaciones que trabajan con grandes bases de datos deben tener acceso rápido a cantidades grandes de datos. Las cachés on-chip son muy pequeñas.	Añadir una caché L3 externa	Pentium III
	Integrar la caché L3 *on-chip*	Pentium 4

Bibliografía

- Arquitectura de Computadoras 3ED, Morris Mano; Editorial Prentice Hall.
- Arquitectura de Computadores, J. Ortega, M. Anguita, A. Prieto; Editorial Paraninfo.