Capítulo 4 Memoria Caché

4.1 Conceptos básicos sobre sistemas de memoria de computadores

Características de los sistemas de memoria

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

**Ubicación**

* **Procesador**
* **Interna(Principal)**
* **Externa(Secundaria)**

**Capacidad**

* **Tamaño de palabra**
* **Numero de palabras**

**Unidad de transferencia**

* **Palabra:** se expresa normalmente en longitudes de 8, 16, 32 bits
* **Bloque:** Conjunto de palabras

**Método de acceso**

* **Acceso secuencial:** La memoria se organiza en unidades de datos llamadas registros, se utiliza un mecanismo de lectura/escritura que debe ir trasladándose desde su posición actual a la deseada, pasando y obviando cada registró intermedio. Ejemplo unidades de Cintas
* **Acceso directo:** La memoria se organiza en vecindades de datos llamados registros, los bloques individuales tienen una dirección única basada en su dirección física, se accede a una vecindad dada y luego se realiza una búsqueda secuencial. Ejemplo unidades de Disco.
* **Acceso aleatorio:** Cada posición de memoria tiene un único mecanismo de acceso cableado físicamente. El tiempo para acceder a una posición dada es constante e independiente de accesos previos. Ejemplo Cache, Memoria DDR4 SDRAM, SRAM.
* **Acceso Asociativo:** Las palabras son recuperadas basándose en una porción de su contenido en lugar de su dirección, el tiempo de recuperación de un dato es constante e independiente accesos previos. Ejemplo Cache.

**Prestaciones**

* **Tiempo de acceso(latencia):** Tiempo que tarda en realizarse una operación de lectura/escritura, tiempo que tarda el mecanismo de lectura/escritura en colocarse en la posición deseada
* **Tiempo de ciclo:** En las RAM es el tiempo de acceso y un tiempo más antes de que pueda iniciarse un segundo acceso a memoria.
* **Velocidad de transferencia:** Velocidad a la que se transfieren los datos a o desde una unidad de memoria, Para las RAM es el inverso del tiempo de ciclo.

Para otras memorias Es TN=TA+(N/R)

TN: Tiempo medio de escritura o de lectura de N bits.

TA: Tiempo de acceso medio.

N: Numero de bits.

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

**Dispositivo físico**

* **Semiconductor:** Memorias semiconductoras
* **Soporte magnético:** Discos y Cintas.
* **Soporte óptico:** l
* **Magneto-óptico:** l

Características físicas

* **Volátil/No volátil:** Si necesita estar alimentada para almacenar los datos
* **Borrable/No borrable:** Si se pueden borrar los datos después de ser grabados

Organización

**JERARQUIA DE MEMORIA:**

En las restricciones de diseño de memoria tenemos:

* ¿Cuánta capacidad?
* ¿Cómo de rápida?
* ¿De qué coste?

La figura 4.1 Ilustra una jerarquía típica. Cuando se desciende en la jerarquía ocurre:



**4.2. PRINCIPIOS BÁSICOS DE LAS MEMORIAS CACHE:**



**4.3. ELEMENTOS DE DISEÑO DE LA CACHE:**



**TAMAÑO DE CACHE:**

Pequeño para que el coste total medio por bit se aproxime al de la memoria principal sola, que sea lo suficientemente grande como para que el tiempo de acceso medio total sea próximo al de la cache sola.

**FUNCIÓN DE CORRESPONDENCIA:** Tenemos:

* **DIRECTA:** Cada bloque de memoria principal corresponde a solo una línea de cache i=jmodulo(m), i:numero de líneas de cache, j:numero de bloques de memoria principal, m: numero de líneas en la cache.
* **ASOCIATIVA:** Cualquier bloque de la memoria principal puede cargarse en cualquier línea de la cache, no es posible en la directa



* **ASOCIATIVA POR CONJUNTOS: Recoge lo positivo de la correspondecia directa y asociativa, sin presentar sus desventajas.**





ALGORITMOS DE SUSTITUCIÓN:

Una vez se ha llenado la cache, para introducir un nuevo bloque debe sustituirse un bloque existente. Para el caso de correspondencia directa, solo hay una posible línea para cada bloque particular y no hay elección posible.

Para las técnicas asociativas deben usarse algoritmos de sustitución, tales algoritmos se implementan en hardware, los más comunes son 4:

1. LRU( Least Recently Used): Se sustituye el bloque que ha sido mantenido en la cache por más tiempo sin haber sido referenciado.
2. FIFO( First In First Out): Puede implementarse fácilmente mediante un buffer circular o Round-Robin.
3. LFU( Least Frecuently Used): Se sustituye el bloque que ha experimentado menos referencias.
4. Aleatoria: Se escoge una línea aleatoriamente y se sustituye.

**POLITICA DE ESCRITURA:**

Tenemos 3:

1. Escritura inmediata: Las operaciones de escritura se hacen tanto en cache como en memoria principal: Las escrituras solo se hacen en cache.
2. Post-escritura:
3. Escritura única:

**TAMAÑO DE LINEA:** Cuando se recupera y ubica un bloque de datos, se recuperan no solo la palabra deseada sino además algunas palabras adyacentes, a medida que aumenta el tamaño de bloque, la tasa de aciertos en principio aumenta y luego disminuye a medida que se hace más grande.

**NUMERO DE CACHES:**

Los sistemas tienen multiples caches, hay 2 aspectos de diseño relacionado:

1. **CACHES MULTINIVEL:** 1 o 2 niveles.
2. **UNIFICADA O PARTIDA(UNIFICADA O SEPARADA):** La cache unificada tiene varias ventajas, tiene una tasa de aciertos mayor, solo se necesita diseñar e implementar una cache, a pesar de las ventajas la tendencia es hacia caches separadas una dedicada a instrucciones y otra a datos.