Departamento de Cs. e Ingeniería de la Computación
Universidad Nacional del Sur

Arquitectura de Computadoras

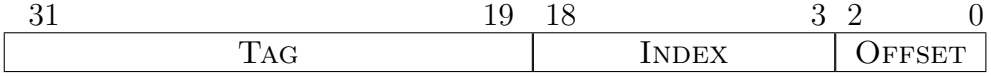
Trabajo Práctico N° 09
Memoria [1]

Primer Cuatrimestre de 2018

Ejercicios

1. Un cierto procesador cuenta con una memoria cache que implementa *mapeo directo* y que tiene tamaño de bloque de ocho bytes.

   a) Dimensionar el tamaño de esta cache en cuanto a número de líneas y a cantidad de memoria destinada a los datos y a los tags. Tener en cuenta que se debe trabajar con paridad por byte para los datos y con paridad única para el tag, y también que hace falta representar el bit de válido.



   b) A partir de la memoria cache bosquejada en el inciso anterior, en caso de adoptar una organización *2-way set associative* que conserve el número de bloques, analizar cuál sería la distribución de los campos de dirección y qué ocurre con el tamaño de memoria para almacenar los tags.

2. Una determinada memoria cache *4-way set associative* consta de 64 líneas, donde la memoria principal contiene 4096 bloques, cada uno de 128 palabras de 32 bits. Mostrar el formato de dirección de memoria principal, en lo que a los campos Tag, Index y Offset respecta.

3. Otra memoria cache con *mapeo directo* tiene 128 líneas, donde la memoria principal contiene 16384 bloques de 16 palabras de 32 bits cada uno. El tiempo de acceso a la cache es de 10ns y el tiempo requerido para llenar un bloque es de 200ns. Cuando una palabra no se encuentra en cache, se traerá primero todo el bloque de memoria principal a cache y luego se transferirá la palabra en cuestión al CPU (es decir, no se está usando la política *load-through*). Inicialmente la cache esta vacía.

   a) Mostrar el formato de la dirección de memoria.

   b) Computar el *hit ratio* para un programa que ejecuta un lazo 10 veces desde la locación 15 a la 200. Nótese que aun cuando la memoria se accede dos veces en un miss, no ocurre un hit para ese caso. Es decir, a los efectos del programa que se este ejecutando sólo se observa una única referencia a memoria.

   c) Computar el tiempo de acceso efectivo para este programa.

[1] Fecha sugerida de finalización de este trabajo práctico: 11 de junio de 2018.

4. Se desea determinar el nivel de asociatividad más conveniente para una cierta arquitectura. Hasta el momento se ha observado el siguiente comportamiento:

   - El *hit time* de un primer nivel de cache *2-way set associative* es de un ciclo de reloj. Debe tenerse en cuenta que esta cache se encuentra en el camino crítico, esto es, su *hit time* determinará la frecuencia del reloj del propio procesador.
   - El organizar esta memoria cache con un mayor grado de asociatividad, puntualmente *4-way set associative* en vez de *2-way set associative*, incrementa el *hit time* en un 10 % (bajo el supuesto de mantener el mismo tamaño de cache).
   - Por otra parte, para una cache de 8Kb de datos, como la que se quiere implementar, el *miss rate* cae de 4.9 % a 4.4 % al incrementar el nivel de asociatividad.

   Asumiendo a manera de simplificación que el *miss time* para resolver un acceso al segundo nivel de cache es de 10 ciclos, y que este segundo nivel de cache tiene un 100 % de *hit rate*:

   a) Comparar el tiempo promedio de acceso bajo ambas organizaciones.

   b) Más allá del resultado a nivel de tiempo de acceso promedio a la cache, ¿qué observación se puede hacer respecto a la perfomance global del sistema?

5. En el contexto de la cache descripta en el ejercicio $1, b)$, y suponiendo que se adopta un algoritmo de reemplazo **LRU**, determinar cuáles de las referencias a memoria de la siguiente sucesión producirán un *miss*, y en tales casos indicar de qué tipo fue (esto es, compulsivo o conflictivo).

01	00B000A0	09	00E000A8
02	00B000A4	10	01E000A0
03	00B000A8	11	01E000A4
04	00B000AC	12	00B000A4
05	00B000B0	13	00B000A8
06	00B000B4	14	00E000A4
07	00B000B8	15	00B000A0
08	00E000A4	16	00B000A4

   Obs: Se denomina "compulsivo" al miss que se produce la primera vez que el CPU accede un bloque de memoria para lectura, mientras que se denomina "conflictivo" al miss que se origina en bloques que han sido previamente reemplazados, como consecuencia del mapeo y no por haber agotado el espacio disponible.

6. Responder a las siguientes preguntas:

   a) ¿Por qué incrementar la asociatividad de una cache puede implicar que el Hit Rate también se incremente?

   b) ¿Por qué incrementar el tamaño de una línea de cache puede implicar que el Hit Rate también se incremente?

   c) ¿Por qué incrementar el tamaño de una línea de cache puede reducir la performance del sistema aun cuando el Hit Rate también se vea incrementado?

   d) ¿Por qué incrementar el tamaño de una línea de podría causar que el Hit Rate disminuya?

7. Considerando que bajo un esquema de dos niveles de memoria cache se verifican las siguientes expresiones:

$$\text{Memory Access Time}_{Average} = \text{Hit Time}_{L1} + \text{Miss rate}_{L1} \times \text{Miss Penalty}_{L1}$$

$$\text{Miss Penalty}_{L1} = \text{Hit Time}_{L2} + \text{Miss Rate}_{L2} \times \text{Miss Penalty}_{L2}$$

$$\begin{aligned}\text{Memory Access Time}_{Average} = \ & \text{Hit Time}_{L1} + \text{Miss Rate}_{L1} \times \\ & \times(\text{Hit Time}_{L2} + \text{Miss Rate}_{L2} \times \text{Miss Penalty}_{L2})\end{aligned}$$

En este contexto, si cada 1000 referencias se observan 100 misses en el primer nivel de cache y 20 misses en el segundo:

a) Determinar los miss rate locales y el miss rate global del segundo nivel.

b) ¿Cual sería el tiempo de acceso (en ciclos) si consideramos que un acceso al primer nivel consume un ciclo, que un acceso al segundo nivel diez ciclos y que el costo del miss del segundo nivel es de cincuenta ciclos.

c) Suponiendo que los datos anteriores corresponden a una memoria cache del segundo nivel con mapeo directo, y que se pasa a una organización *2-way set associative*, lo cual reduce el miss rate en un 25 % y que para no comprometer el período del reloj se incrementa a once ciclos el acceso al segundo nivel de cache, ¿cuál será la variación en el miss penalty en el primer nivel con este cambio de organización del segundo nivel?

Obs: El *miss rate local* es el número de misses de esa cache dividido por el número total de accesos a la misma. Análogamente, el *miss rate global* es el número de misses de esa cache dividido por el número total de accesos a memoria generados por el CPU. Nótese que para dos niveles de cache sería igual a $\text{Miss Rate}_{L1} \times \text{Miss Rate}_{L2}$.

8. La jerarquía de memoria de una cierta arquitectura cuyo tamaño de palabra es de 32 bits presenta las siguientes características:

- Block size = 1 word
- Memory bus width = 1 word
- Miss rate = 3 %
- Memory accesses per instruction = 1.2
- Cache miss penalty = 32 cycles
- Average cycles per instruction (ignoring cache miss) = 2

Por otra parte, esta arquitectura organiza la comunicación con memoria adoptando los siguientes retardos:

- 4 ciclos de reloj para enviar la dirección.
- 24 ciclos de reloj para acceder cada palabra.
- 4 ciclos de reloj para enviar un dato de una palabra.

En este contexto, se desean comparar las alternativas de interleaving contra el empleo de memorias (y buses) más anchos. Se debe tener en cuenta que si se cambia el tamaño del bloque a dos palabras, el miss rate cae a un 2 % y que para el caso de un bloque de 4 palabras el miss rate es de un 1 %. Realizar los cómputos comparativos usando como métrica el CPI, dado que se asume que los cambios introducidos no afectarán ni al tiempo de ciclo ni al número total de instrucciones.

9. Las memoria de tipo DDR (*Double Data Rate*), esto es, que transfieren en ambos flancos del pulso de reloj, pueden ser fácilmente identificadas gracias a la forma adoptada para especificar los chips de memoria del tipo SDRAM (*Syncronous Dinamic RAM*), puesto que usualmente se indica la cantidad de bits que pueden ser accedidos por unidad de tiempo. Por ejemplo, con un bus de 133Mhz se alude a una DDR266; con uno de 200Mhz a una DDR400, y así sucesivamente.

   Desafortunadamente, también existe otra forma de especificar las características de los módulos de memoria, lo cual en ocasiones acarrea algunas complicaciones. Estamos haciendo referencia a las memorias DIMM (*Dual Inline Memory Module*), las cuales contabilizan la velocidad a nivel de bytes por segundo. Por caso, recordando que en el bus del sistema se transfieren bytes en paralelo, para un DDR300 (esto es, reloj del bus a 150Mhz), corresponderá una denominación DIMM de PC2400 (ya que el módulo puede acceder a 64bits en paralelo, esto es, $2 \times 150 \times 8 = 2400$), en otras palabras, una transferencia pico de 2400Mb por segundo.

   Teniendo en cuenta que las formas alternativas de especificar las características de los módulos de memoria antes introducidas resultan aplicables tanto a las las memorias DDR, DDR2 y DDR3, determinar los nombres comerciales y la frecuencia de reloj que corresponden a cada uno de los siguientes módulos de memoria:

   a) Un módulo DIMM tipo DDR que alcanza un pico de transferencia de 3200Mb por segundo.

   b) Un módulo DIMM tipo DDR2 que alcanza picos de transferendia de 6400b por segundo.

   c) Un módulo DIMM tipo DDR3 que alcanza picos de transferendia de 12800Mb por segundo.

Referencias

[HP96] Hennessy, J., and Patterson, D. *Computer Architecture*, second ed. Morgan Kaufmann, 1996.

[HP06] Hennessy, J., and Patterson, D. *Computer Architecture*, fourth ed. Morgan Kaufmann, 2006.