EJERCICIOS - ENUNCIADOS

ARQUITECTURA DE COMPUTADORES (GII-AC)

BLOQUE 2: DISEÑO DE LA JERARQUÍA DE MEMORIA

Tema 2-2: EJERCICIOS MEMORIA PRINCIPAL

Problema 2-2-1 (HP96, example, pp.432). Considerar la siguiente descripción de un computador que dispone de Procesador, Memoria Cache Unificada y Memoria Principal.

Si el procesador dispusiera de una memoria cache perfecta (nunca falla), ejecutaría un programa benchmark y obtendría los siguientes resultados de prestaciones:

Número promedio de ciclos por instrucción (ignorando fallos de cache): 2

Accesos a memoria por instrucción: 1.2

Sin embargo, la memoria cache real tiene las siguientes características:

Tamaño de bloque/línea: 1 palabra de 32 bits

Frecuencia de fallos: 3%

Penalización por fallo de cache (Caso 0): 32 ciclos

Si se cambia el tamaño de bloque de la memoria cache a 2 palabras (Casos 1, 2, 3), la frecuencia de fallos disminuye al 2%, y si dispusiera de bloques de 4 palabras (Casos 4, 5, 6, 7), la frecuencia de fallos disminuye a 1%.

La memoria principal DRAM tiene las siguientes características:

Ancho del bus FSB procesador-memoria: 1 palabra

Duración del envío de dirección: 4 ciclos de reloj

Duración del tiempo de acceso por palabra y bloque de memoria DRAM: 24 ciclos

Tiempo para enviar al procesador una palabra desde la memoria principal: 4 ciclos

¿Cuál es la mejora en prestaciones respecto al computador de referencia con cache de 1 palabra/bloque cuando se aumenta el tamaño de bloque a 2 y 4 palabras y se utiliza tanto memoria simple como memoria entrelazada de 2 ó 4 vías? Considerar que el ancho de bus FSB coincide con el número de palabras que aloja cada módulo de memoria DRAM en una dirección.

Solución:

Lo que nos pide este problema es que realicemos un análisis comparativo de las prestaciones de una máquina cuando: **variamos el tamaño de bloque de la cache de 2 a 4 palabras**, **aumentamos el ancho**

del bus a memoria de 1 a 2 palabras, y **utilizamos una memoria entrelazada de 2 o 4 vías** frente a utilizar una memoria no entrelazada, todo ello con respecto al sistema base de referencia cuyas características aparecen en el enunciado.

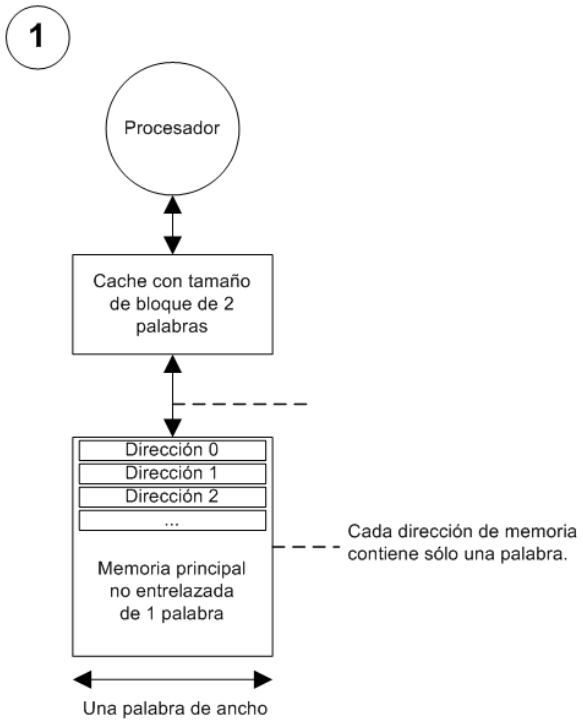
Dichas características nos informan que, en la **máquina de referencia**, en promedio aparece un **20% de cargas y almacenamientos** en los programas. Puesto que los accesos a memoria por instrucción son de 1.2, como toda instrucción accede a memoria, quedaría un 0.2 (es decir, un 20%) de accesos que serían de cargas y almacenamientos. Además, en la máquina de referencia la **tasa de fallos de la memoria cache es de un 3%**; es decir, de cada 100 accesos a la cache, 3 de esos accesos requieren que se vaya a buscar el dato a la memoria principal, con la consecuente **penalización de 32 ciclos**.

Hay que resaltar que en este problema no estamos diferenciando entre varios niveles de cache, sino que consideramos la cache en su conjunto y nos centramos en estudiar su relación con la memoria principal. Por tanto, las tasas de fallos que **se proporcionan son las tasas de fallos globales de la cache**; es decir, el porcentaje de veces que la cache falla y tiene que acceder a la memoria principal para obtener el dato solicitado.

Las situaciones que vamos a estudiar de acuerdo con los datos proporcionados por el problema son las siguientes:

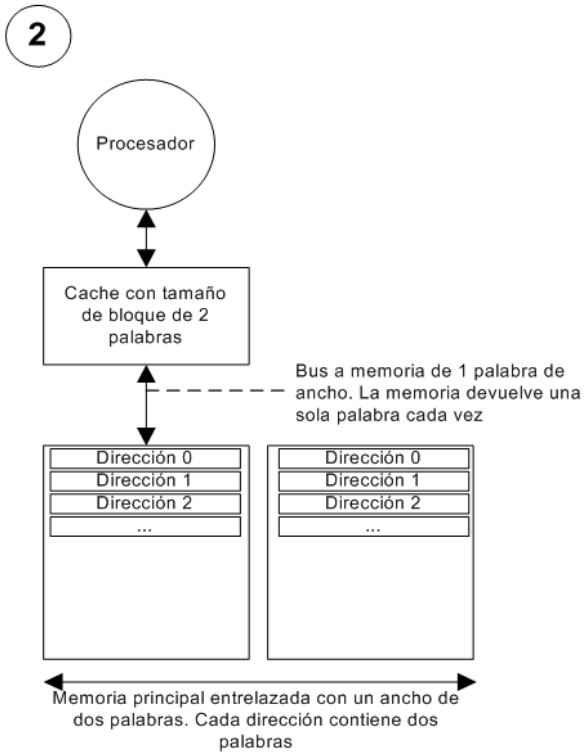
- **Caso 1**: Máquina con cache con tamaño de bloque de dos palabras, ancho de bus de la memoria principal de 1 palabra y memoria principal no entrelazada de ancho de una palabra. En este caso, para guardar un bloque de cache completo desde la memoria principal, hay que hacer **dos accesos completos a memoria**, ya que la memoria sólo permite direccionar una palabra a la vez y el bus tiene un ancho de una sola palabra.

Número de ciclos de penalización= 2 × 4 (envío dirección a memoria DRAM) + 2 × 24 (latencia DRAM) + 2 × 4 (envío datos a cache) = 64 ciclos



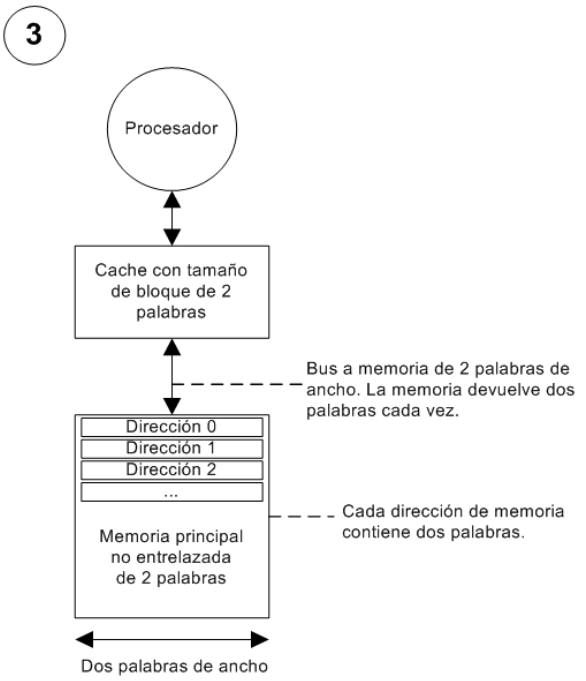
- **Caso 2**: Máquina con cache con tamaño de bloque de 2 palabras, ancho de bus de la memoria principal de 1 palabra y memoria principal entrelazada de 2 vías, siendo el ancho de cada vía de 1 palabra. En este caso, al tener una memoria entrelazada de dos vías, podemos direccionar las dos palabras simultáneamente, pero hay que transmitirlas en serie, ya que el ancho del bus es de una sola palabra. Por tanto, hay que enviar la dirección una sola vez y acceder una sola vez a la dirección de memoria, pero hay que realizar dos envíos (uno por palabra) desde la memoria principal a la cache.

Número de ciclos de penalización= 1 × 4 (envío dirección a memoria DRAM) + 1 × 24 (latencia DRAM) + 2 × 4 (envío datos a cache) = 36 ciclos



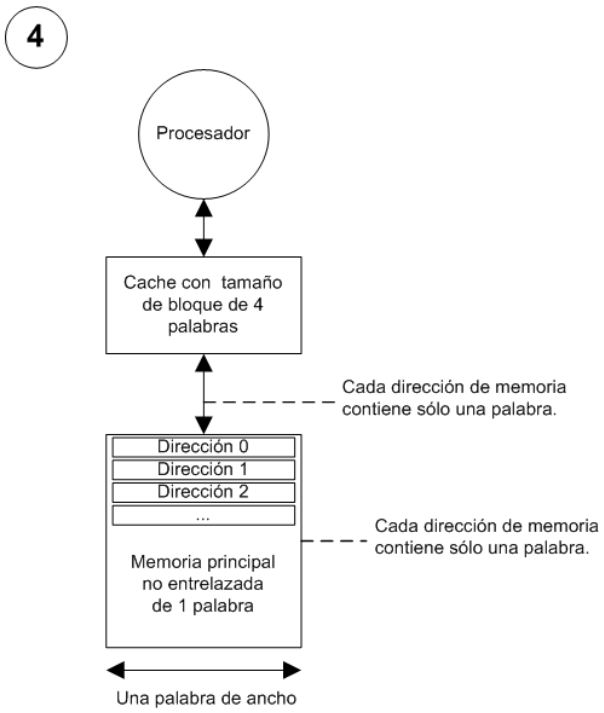
- **Caso 3**: Máquina con cache con tamaño de bloque de 2 palabras, ancho de bus de la memoria principal de 2 palabras y memoria principal no entrelazada de ancho de dos palabras. En este caso, hay que realizar **un solo acceso completo** a la memoria principal por cada fallo de la cache, ya que cada dirección de memoria contiene dos palabras y es posible enviarlas por el bus ya que su ancho se ha aumentado a dos palabras.

Número de ciclos de penalización= 1 × 4 (envío dirección a memoria DRAM) + 1 × 24 (latencia DRAM) + 1 × 4 (envío datos a cache) = 32 ciclos



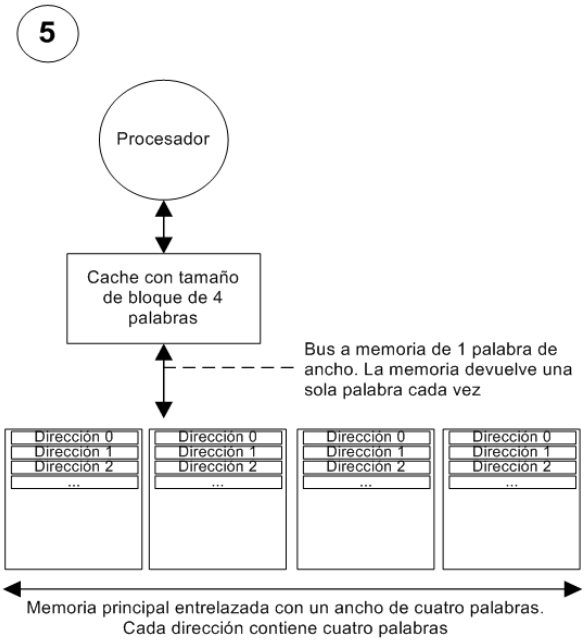
- **Caso 4**: Máquina con cache cuyo tamaño de bloque es de 4 palabras, ancho de bus de la memoria principal de 1 palabra y memoria principal no entrelazada de ancho de una palabra. En este caso, por cada fallo de la cache habría que realizar **cuatro accesos completos** a memoria, ya que en cada acceso la memoria sólo nos proporciona una palabra y el bus sólo puede transmitir una palabra.

Número de ciclos de penalización= 4 × 4 (envío dirección a memoria DRAM) + 4 × 24 (latencia DRAM) + 4 × 4 (envío datos a cache) = 128 ciclos



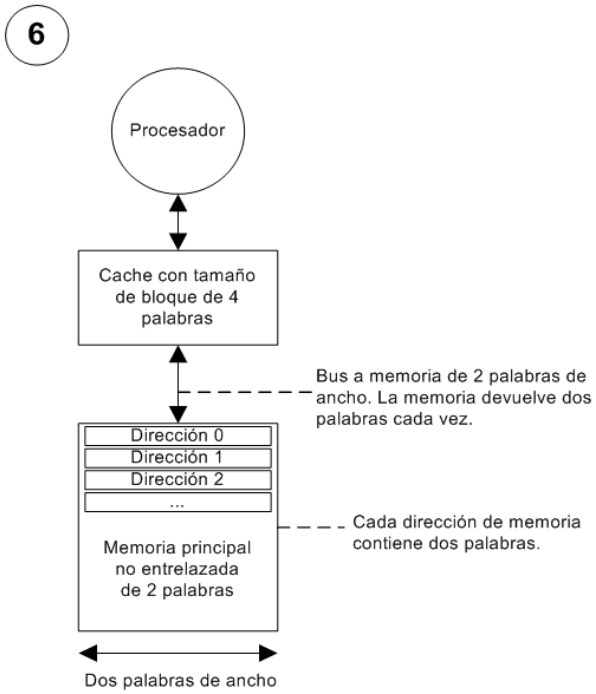
- **Caso 5**: Máquina con cache cuyo tamaño de bloque es de 4 palabras, ancho de bus de la memoria principal de 1 palabra y memoria principal entrelazada de 4 vías, siendo el ancho de cada vía de una palabra. En este caso, como en el Caso 2, podemos direccionar las cuatro palabras de una sola vez, pero necesitamos que se realicen cuatro envíos entre la memoria principal y la cache para obtener las cuatro palabras, ya que la capacidad del bus está limitada a una palabra. Por tanto, habría que **direccionar una vez, acceder una vez a la memoria y realizar cuatro envíos**.

Número de ciclos de penalización= 1 × 4 (envío dirección a memoria DRAM) + 1 × 24 (latencia DRAM) + 4 × 4 (envío datos a cache) = 44 ciclos



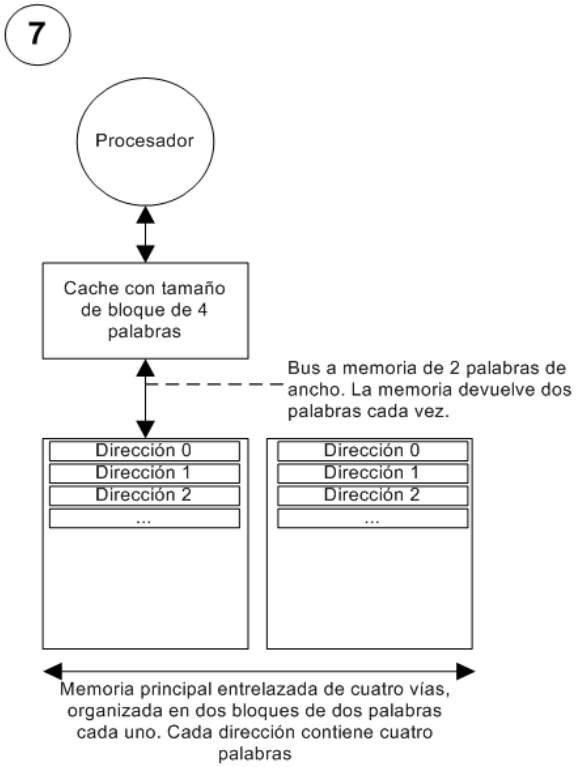
- **Caso 6**: Máquina con cache cuyo tamaño de bloque de 4 palabras, ancho de bus de la memoria principal de 2 palabras y la memoria principal es no entrelazada de ancho de dos palabras. En este caso, serán necesarios **2 accesos completos** a memoria para obtener el bloque de cache. En cada acceso obtendríamos dos palabras.

Número de ciclos de penalización= 1 × 4 (envío dirección a memoria DRAM) + 1 × 24 (latencia DRAM) + 4 × 4 (envío datos a cache) = 44 ciclos



- **Caso 7**: Máquina con cache cuyo tamaño de bloque es de 4 palabras, ancho de bus de la memoria principal de 2 palabras y memoria principal entrelazada de 2 vías, siendo el ancho de cada vía de dos palabras. En este caso, al tener una memoria entrelazada de 2 vías con un ancho de 2 palabras en cada vía, podemos obtener con un solo direccionamiento las cuatro palabras del bloque y, al disponer de un bus de dos palabras, sólo necesitamos realizar dos envíos

Número de ciclos de penalización= 1 × 4 (envío dirección a memoria DRAM) + 1 × 24 (latencia DRAM) + 2 × 4 (envío datos a cache) = 36 ciclos



En base a lo explicado en los párrafos anteriores, en la siguiente Tabla 1 se resume el número de ciclos que emplearía cada una las configuraciones en obtener un bloque de cache, asumiendo que el envío de la dirección tarda 4 ciclos, la latencia de la memoria DRAM por palabra es 24 ciclos y el envío de una palabra al procesador a través del bus FSB es de 4 ciclos.

Tabla 1. Obtención de los ciclos de penalización por fallo de la memoria cache.

Configuración	Nº de envíos de la dirección = Nº total de accesos	Nº de envíos de palabra al procesador	Nº de ciclos (cálculo)	Nº de ciclos (resultado)
1	2	2	$4 \times 2 + 2 \times 24 + 2 \times 4$	64
2	1	2	$4 \times 1 + 1 \times 24 + 2 \times 4$	36
3	1	1	$4 \times 1 + 1 \times 24 + 1 \times 4$	32
4	4	4	$4 \times 4 + 4 \times 24 + 4 \times 4$	128
5	1	4	$4 \times 1 + 1 \times 24 + 4 \times 4$	44
6	2	2	$4 \times 2 + 2 \times 24 + 2 \times 4$	64
7	1	2	$4 \times 1 + 1 \times 24 + 2 \times 4$	36

Para medir las mejoras en las prestaciones de cada una de las configuraciones con respecto al sistema de referencia, calcularemos el CPI. Para ello, partiremos del CPI base que nos proporciona el problema y le sumaremos las penalizaciones debidas a la cache para obtener el CPI final. Estas penalizaciones (en ciclos por instrucción) serán debidas a los accesos por instrucción, el porcentaje de fallos que se produce en dichos accesos, y la penalización en ciclos por cada fallo. Formalmente:

$$\mathbf{CPI = CPI_{BASE} + Penalización/Fallo \times Fallos/Acceso \times Accesos/Instrucción}$$

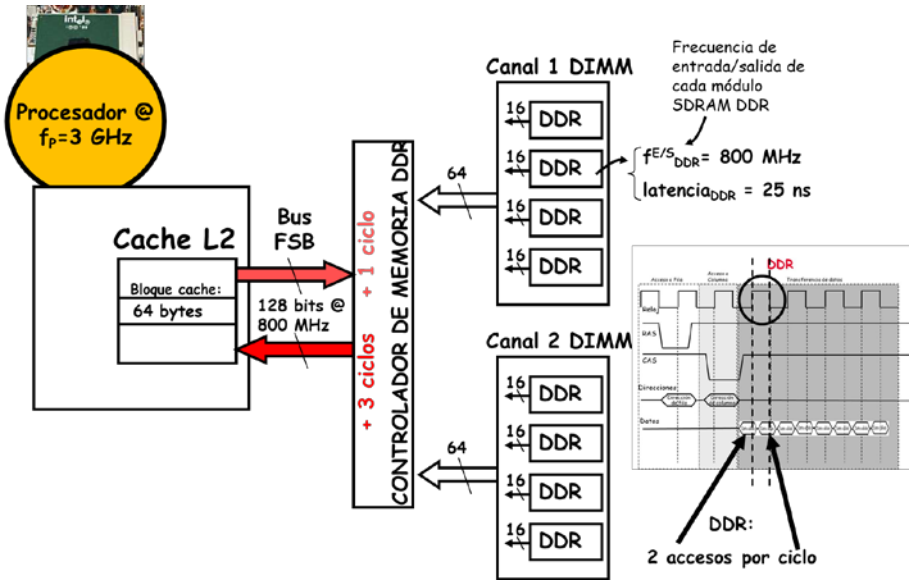
En la siguiente Tabla 2 se muestra el cálculo del CPI tanto para el sistema de referencia como para las diferentes configuraciones (casos). Se han considerado los fallos de cache indicados en el enunciado, y los ciclos de penalización por fallos los calculados en la Tabla 1.

Tabla 2. Obtención del CPI de los distintos casos indicados en la Tabla 1.

Caso	CPI_{BASE}	Accesos/ Instrucción	Fallos/Acceso (%)	Penalizaciones/Fallo [ciclos]	CPI	Speed-Up
0-Ref.	2.0	1.2	3%	32	3.15	1.00X
1	2.0	1.2	2%	64	3.54	0.89X
2	2.0	1.2	2%	36	2.86	1.10X
3	2.0	1.2	2%	32	2.77	1.14X
4	2.0	1.2	1%	128	3.54	0.89X
5	2.0	1.2	1%	44	2.53	1.25X
6	2.0	1.2	1%	64	2.77	1.14X
7	2.0	1.2	1%	36	2.43	1.30X

Como puede verse en esta Tabla 2, **la mejor** de las configuraciones que es la que proporciona un CPI más bajo, corresponde a **la configuración 7 con un CPI de 2.43** (memoria cache con tamaño de bloque de 4 palabras, memoria principal entrelazada de 2 vías con dos palabras por vía y bus FSB con un ancho de dos palabras). **La segunda mejor** configuración sería **la 5 con un CPI de 2.53** (memoria cache con tamaño de bloque de 4 palabras, memoria principal entrelazada de 4 vías con una palabra por vía y bus a memoria principal de una palabra). Por último, **en el tercer puesto hay un empate** entre las **configuraciones 6** (memoria cache con tamaño de bloque de 4 palabras, memoria principal no entrelazada con un ancho de dos palabras, y bus a memoria principal de dos palabras) **y la 3** (memoria cache con tamaño de bloque de 2 palabras, memoria principal no entrelazada con un ancho de dos palabras y bus a memoria principal de dos palabras), ambas **con un CPI de 2.77.**

Problema 2-2-2. Se desea analizar la penalización del CPI de un programa debido a los accesos a la memoria principal utilizando un sistema computador como el que se muestra a continuación con dos canales independientes de acceso a memoria principal (Canales 1 y 2).



Por simulación se sabe que, si el programa se ejecuta sin fallos de la cache L2, el CPI sería: CPI_{ideal}=0.3 ciclos del procesador (f_P= 3 GHz); y la frecuencia de fallos de la cache L2 por instrucción retirada es de FF=5%. Por otro lado, considerar que:

- Los accesos a la memoria principal se originan por fallos de cache L2 cuyo tamaño de bloque es de 64 bytes.
- El bus FSB tiene un $ancho_{FSB}$= 128 bits (16 bytes), y una frecuencia de f_{FSB}= 800 MHz (periodo: T_{FSB} = 1.25 ns).
- El envío de la dirección del bloque de cache L2 que falla al controlador de memoria DDR tarda 4 ciclos de bus FSB (envío dirección desde L2 a controlador memoria SDRAM DDR: 4 x T_{FSB} = 5 ns).
- El envío de la dirección del bloque de L2 desde el controlador a los módulos DIMM de memoria principal tarda 1 ciclo del bus FSB. Suponer que el bloque de L2 está distribuido homogéneamente de forma entrelazada entre todos los chips DDR, y que el controlador de memoria sólo puede resolver un fallo de L2 en cada instante; es decir, que no acepta un nuevo acceso a memoria principal hasta que no haya finalizado el anterior.
- La latencia de cada uno de los chips DDR que forman los módulos DIMM es de $latencia_{DDR}$=25 ns (tiempo que transcurre desde que un chip DDR recibe la dirección hasta que proporciona -lectura- o almacena -escritura- los 16 bit (2 bytes) a través de su puerto de datos), y su frecuencia de entrada/salida es de $f^{E/S}_{DDR}$= 800 MHz.
- El retraso de enviar los datos desde el controlador de memoria a la cache L2 es de 3 ciclos del bus FSB.

Calcular:

(a) El ritmo de transferencia máximo BW_{DDR} [bytes/segundo] que exhibe cada chip de memoria DDR.
(b) El ritmo de transferencia en cada acceso a memoria que es exhibido por el bus FSB: BW_{FSB}.
(c) Los ciclos de penalización por cada fallo de cache L2.
(d) El CPI real que aparecería cuando se consideran los fallos de la cache L2.
(e) Porcentaje del CPI debido a la penalización por transferencia del bloque de cache por el bus FSB.
(f) Porcentaje del CPI debido a la penalización por la latencia de los módulos DDR.

Solución:

(a) $BW_{DDR} = f_{DDR}^{E/S} \times Bytes = 800MHz \times 2\ Bytes \times 2\ envíos_{DDR} / ciclo_{DDR} = 3200MB/s = 3.2GB/s$

(b) $BW_{FSB} = f_{FSB} \times Bytes = 800MHz \times 16\ Bytes = 12800MB/s = 12.8GB/s$

(c) El ritmo de transferencia de todos los módulos DDR en paralelo (3.2GB/s × 8DDR = 25.6GB/s) es superior al del FSB (12.8GB/s), por lo que el número de ciclos que se tarda en transferir un bloque de L2 viene determinado por el FSB.

A	Envío de la dirección desde L2 al controlador de memoria	4 ciclos
B	Envío de la dirección desde el controlador de memoria al módulo DIMM	1 ciclo
C	Latencia de los circuitos de memoria DDR	$\frac{25ns}{\frac{1}{800MHz}} = 20ciclos@800MHz$
D	Envío del bloque desde DDR a L2 (pipeline con el controlador de memoria)	$\frac{64bytes/L2}{16bytes/ciclo_{FSB}} = 4ciclos@800MHz$
E	Retraso en enviar los primeros datos desde el controlador a la L2	3 ciclos
	PENALIZACIÓN$_{L2}$: P $_{L2}$ = A+B+C+D+E =	32 ciclos@800MHz = 120 ciclos @ 3 GHz

(d) $CPI = CPI_{ideal} + FF \times P_{L2} = \frac{1}{3} + 0.05 \times 120 = 6.333$ ciclos (del procesador, 3 GHz)

(e) De los 32 ciclos del bus FSB de la penalización de un fallo de L2, la originada por la transferencia del bloque de cache L2 es (ver apartado c): 4 ciclos en la frecuencia FSB o (4/32)×120 en ciclos del procesador.

$$\frac{CPI_{FSB}}{CPI} = \frac{FF \times P_{L2}\big|_{FSB}}{CPI} = \frac{0.05 \times 4 \times \frac{120}{32}}{6.333} = 11.8\%$$

(f) La penalización originada por la latencia del módulo DDR es (ver apartado c): 20 ciclos en la frecuencia FSB o 20×120/32 en ciclos del procesador.

$$\frac{CPI_{LatenciaDDR}}{CPI} = \frac{FF \times P_{L2}\big|_{LatenciaDDR}}{CPI} = \frac{0.05 \times 20 \times \frac{120}{32}}{6.333} = 59.2\%$$