Arquitectura e Ingeniería de Computadores - 3er Curso. Sol. del 2º Parcial 10/1/2022

Responde cada pregunta en una hoja distinta. Tiempo disponible: 2h30m

1. **(3 puntos)** Se dispone de un procesador MIPS con ejecución fuera de orden y especulación hardware basada en el algoritmo de Tomasulo. Las instrucciones atraviesan las siguientes etapas: IF (búsqueda de instrucciones), I (decodificación y lanzamiento de las instrucciones), En (ejecución en el operador multiciclo correspondiente), WB (escritura en el bus común de datos) y C (confirmación de las instrucciones). El ROB dispone de 32 entradas numeradas de la #0 a la #31. El procesador dispone de un predictor de saltos del tipo *Branch Target Buffer* (BTB) de 1 bit que ofrece la predicción al final de la etapa IF.

   Se pretende evaluar el comportamiento del procesador ante el siguiente bucle que multiplica las componentes de dos vectores de 10 elementos:

```
     .data
x: .double 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
y: .double 2, 2, 2, 2, 2, 2, 2, 2, 2, 2
z: .double 0, 0, 0, 0, 0, 0, 0, 0, 0, 0

     .text 0x000000
start:
    dadd r1, r0, x
    dadd r2, r0, y
    dadd r3, r0, z
    dadd r4, r1, 80 ; 10 Elements * 8
    dadd r8, r0, 8
loop:
    l.d f1, 0(r1) ; Load X
    l.d f2, 0(r2) ; Load Y
    mul.d f4, f1, f2 ; X*Y
    s.d f4, 0(r3) ; Store Z
    dadd r1, r1, r8
    dadd r2, r2, r8
    dadd r3, r3, r8
    dsub r5, r4, r1
    bnez r5, loop
    dadd r8, r0, r0
    trap 0
```

   La figura siguiente muestra el diagrama instrucciones–tiempo correspondiente a la primera iteración del bucle:

PC	Instruccion	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
start	dadd r1,r0,x	IF	I	E1	WB	C																								
4100	dadd r2,r0,y		IF	I	E1	WB	C																							
4104	dadd r3,r0,z			IF	I	E1	WB	C																						
4108	dadd r4,r1,80				IF	I	E1	WB	C																				
4112	dadd r8,r0,8					IF	I	E1	WB	C																				
loop	**l.d f1,0(r1)**						IF	I	AC	L1	L2	WB	C																	
4120	l.d f2,0(r2)							IF	I	AC	-	L1	L2	WB	C															
4124	mul.d f4,f1,f2								IF	I	-	-	-	-	M1	M2	M3	M4	M5	M6	M7	WB	C							
4128	s.d f4,0(r3)									IF	I	AC	-	-	-	-	-	-	-	-	-	-	-	C	L1	L2				
4132	dadd r1,r1,r8										IF	I	E1	-	WB	-	-	-	-	-	-	-	-	-	C					
4136	dadd r2,r2,r8											IF	I	E1	-	WB	-	-	-	-	-	-	-	-	-	C				
4140	dadd r3,r3,r8												IF	I	E1	-	WB	-	-	-	-	-	-	-	-	-	C			
4144	dsub r5,r4,r1													IF	I	E1	-	WB	-	-	-	-	-	-	-	-	-	C		
4148	bnez r5,loop														IF	I	-	-	E1	WB	-	-	-	-	-	-	-	-	C	
4152	dadd r8,r0,r0															IF	I	E1	WB	-	-	-	-	-	-	-	-	-	x	
4156	trap 0																IF	I	-	-	-	-	-	-	-	-	-	-	x	
	<nop>																	if	if	if	if	if	if	if	if	if	if	if	X	
loop	l.d f1,0(r1)																													IF

   Se solicita:

   a) Considera la instrucción `mul.d f4, f1, f2` e indica el ciclo de reloj en el cual:

      1) se copia el valor del registro fuente f1 a la estación de reserva.

2) se copia el valor del registro fuente f2 a la estación de reserva.
3) se libera la estación de reserva utilizada.
4) se almacena el resultado de la ejecución en el ROB.
5) se almacena el resultado en el registro destino f4.

b) Suponiendo, para simplificar, que el ROB, las estaciones de reserva y los operadores estaban vacíos al comenzar la ejecución código suministrado, indica los valores y las marcas de los registros f1, f2, r1 y r5 en el ciclo de reloj en el que la instrucción `mul.d f4,f1,f2` realiza la fase C. Asume que los registros no inicializados almacenan el valor 0 inicialmente.

c) ¿Cuál será el número de ciclos consumido por una iteración cuando el predictor acierta? ¿Y cuando falla?

d) Indica las siguientes características del operador de carga/almacenamiento utilizado:

1) Latencia del operador.
2) ¿Está el operador segmentado?
3) Número mínimo de buffers de lectura necesarios para ejecutar todas las iteraciones del bucle sin interrumpir en ningún momento el lanzamiento a ejecución de las instrucciones de carga.

e) ¿Cómo obtiene la instrucción de salto el valor de r5 y en qué ciclo lo hace?

Justifica en todos los casos tus respuestas con detalle.

Solución:

a) Considera la instrucción `mul.d f4, f1, f2` e indica el ciclo de reloj en el cual:

1) se copia el valor del registro fuente f1 a la estación de reserva: ciclo 11, cuando la instrucción `l.d f1, 0(R1)` escribe su resultado en el bus de datos común (etapa WB)
2) se copia el valor del registro fuente f2 a la estación de reserva: ciclo 13, cuando la instrucción `l.d f1, 0(R1)` escribe su resultado en el bus de datos común (etapa WB)
3) se libera la estación de reserva utilizada: cuando la multiplicación termina su ejecución y escribe su resultado en el ROB (etapa WB, ciclo 21)
4) se almacena el resultado de la ejecución en el ROB: cuando la multiplicación termina su ejecución y escribe su resultado en el ROB (etapa WB, ciclo 21)
5) se almacena el resultado en el registro destino f4: cuando la ejecución de la instrucción es confirmada (etapa C, ciclo 22)

b) Suponiendo, para simplificar, que el ROB, las estaciones de reserva y los operadores estaban vacíos al comenzar la ejecución del bucle, indica los valores y las marcas de los registros f1, f2, r1 y r5 en el ciclo de reloj en el que la instrucción `mul.d f4,f1,f2` realiza la fase C. Asume que los registros no inicializados almacenan el valor 0 inicialmente, y que la primera entrada del ROB se identifica como #0.

- f1: 1 (marca: –)
- f2: 2 (marca: –)
- r1: x (marca: #9)
- r5: 0 (marca:#12)

c) ¿Cuál será el número de ciclos consumido por una iteración cuando el predictor acierta? 9 ciclos ¿Y cuando falla? 23 ciclos

d) Indica las siguientes características del operador de carga/almacenamiento utilizado:

1) Latencia del operador.: 2 ciclos
2) ¿Está el operador segmentado? No, ya que si lo estuviera la etapa L1 de la segunda carga (`l.d f2, 0(R2)`) se realizaría en el ciclo 10 y no en el 11.

3) Número mínimo de buffers de lectura necesarios para ejecutar todas las iteraciones del bucle sin necesidad de introducir ciclos de parada en la etapa Issue: 2 buffers de lectura serían suficiente puesto que de una iteración a la siguiente no existe solapamiento en la ejecución de las instrucciones de carga existentes en el bucle.

e) ¿Cómo obtiene la instrucción de salto el valor de r5 y en qué ciclo lo hace? Lo lee del bus de datos común cuando lo escribe instrucción anterior durante su etapa WB (ciclo 17).

□

2. (3 **puntos)** Se propone ejecutar el código de multiplicación de vectores del ejercicio anterior en un procesador superescalar con arquitectura MIPS, y ejecución fuera de orden. Este procesador superescalar dispone de 4 vías y de los siguientes operadores:

	Nº Operadores	Latencia	Características
Carga/Almacenamiento	4	2	No segmentada; 8 buffers de lectura y 8 de escritura
Suma/Resta CF	4	2	Segmentada; 8 estaciones de reserva
Multiplicador/Divisor CF	4	4	Segmentada; 8 estaciones de reserva
Enteros/Saltos	4	1	8 estaciones de reserva

El procesador emplea un predictor de 2 bits con histéresis para resolver los riesgos de control, generando la predicción en la etapa IF. El ROB dispone de 32 entradas numeradas de la #0 a la #31.

a) Completa la tabla adjunta con el diagrama de instrucciones/tiempo partiendo del inicio de una iteración intermedia del bucle (desde *loop*) hasta que se busquen 20 instrucciones. Considera que el predictor acierta y que al inicio de la iteración el ROB y las estaciones de reserva están vacías, y que los registros no tienen marcas.

b) Muestra el estado del ROB al final del ciclo 4.

Solución:

a) Diagrama instrucciones/tiempo

Num.	Instruccion	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	dadd r8,r0,#8	IF	X															
2	l.d f1,0(r1)	IF	I	AC	L1	L2	WB	C										
3	l.d f2,0(r2)	IF	I	AC	L1	L2	WB	C										
4	mul.d f4,f1,f2	IF	I					M1	M2	M3	M4	WB	C					
5	s.d f4,0(r3)		IF	I	AC								C	L1	L2			
6	dadd r1,r1,r8		IF	I	E1	WB							C					
7	dadd r2,r2,r8		IF	I	E1	WB							C					
8	dadd r3,r3,r8		IF	I	E1	WB								C				
9	dsub r5,r4,r1			IF	I		E1	WB						C				
10	bnez r5,loop			IF	I				E1	WB				C				
11	dadd r8,r0,r0			IF	X													
12	trap 0			IF	X													
13	dadd r8,r0,#8				IF	X												
14	l.d f1,0(r1)				IF	I	AC	L1	L2	WB				C				
15	l.d f2,0(r2)				IF	I	AC	L1	L2	WB					C			
16	mul.d f4,f1,f2				IF	I					M1	M2	M3	M4	WB	C		
17	s.d f4,0(r3)					IF	I	AC								C	L1	L2
18	dadd r1,r1,r8					IF	I	E1	WB							C		
19	dadd r2,r2,r8					IF	I	E1	WB							C		
20	dadd r3,r3,r8					IF	I	E1	WB								C	

b) Estado del ROB al final del ciclo 4.

ROB						
#	Busy	Instrucción	Destino	Valor	Completado	Predicción
0	SI	l.d f1,0(r1)	f1		NO	
1	SI	l.d f2,0(r2)	f2		NO	
2	SI	mul.d f4,f1,f2	f4		NO	
3	SI	s.d f4,0(r3)	s1		NO	
4	SI	dadd r1,r1,r8	r1		NO	
5	SI	dadd r2,r2,r8	r2		NO	
6	SI	dadd r3,r3,r8	r3		NO	
7	SI	dsub r5,r4,r1	r5		NO	
8	SI	bnez r5, loop	loop		NO	Salta

sksksks

□

3. **(2 puntos)** Un procesador A a 4.8GHz incorpora una cache L2 de 256KB y una cache **L3 inclusiva** de 1,5 MB. El fabricante cambia el diseño de la jerarquía de cache. En concreto, el procesador B aumenta a 512KB la cache L2 y reduce a 1,1MB la cache L3, que pasa a ser **exclusiva**.

   El tiempo de acierto es de 8 y 30 ciclos en el procesador A, y de 10 y 25 ciclos en el procesador B, para las caches L2 y L3, respectivamente. La tasa de fallos es 50 % y 40 % en el procesador A y de 40 % y 50 % en el procesador B para las caches L2 y L3, respectivamente. El tiempo medio de acceso a la memoria principal es, en ambos casos, 130 ciclos del procesador. En ambos procesadores la $TF_{L1} = 0,10$.

   Se pide:

   a) Indica una ventaja y una desventaja de las caches exclusivas frente a las inclusivas.

   b) Calcula la **capacidad efectiva** (correspondiente a bloques almacenados distintos) de almacenamiento en MB conjunta L2+L3 en el procesador A y el procesador B.

   c) Calcula la PF de L1 para el procesador A.

   d) Razona cuál de los dos procesadores realiza más accesos a memoria principal sabiendo que en ambos casos el número de accesos a la cache L1 es el mismo.

   Solución:

   a) Indica una ventaja y una desventaja de las caches exclusivas.

   Ventaja: aprovechan mejor el espacio al no disponer de bloques replicados en ambos niveles.
   Desventaja: complican la implementación de la coherencia.

   b) Calcula la **capacidad efectiva** (correspondiente a bloques almacenados distintos) de almacenamiento en MB conjunta L2+L3 en el procesador A y el procesador B.

   Cache inclusiva : 256KB + 1,5 MB - 256KB = 1,5 MB
   Cache exclusiva: 512KB + 1,1MB = 1,612MB

   c) Calcula la PF de fallo de L1 para el procesador A.

$$PF_{L1} = TA_{L2} + TF_{L2} \times (TA_{L3} + TF_{L3} \times PF_{L3}) =$$

$$= 8 + 0,5 \times (30 + 0,4 \times 130) = 49 \text{ ciclos}$$

   d) El número de accesos a memoria viene dado por el número de accesos a L1 (que en ambos casos es el mismo) multiplicado por la tasa de fallos global:

$$TF_{L1} \times TF_{L2} \times TF_{L3}$$

   Como el producto en ambos procesadores coincide:

$$(\text{Proc A})TF_{L1} \times 0{,}5 \times 0{,}4 = (\text{Proc B})TF_{L1} \times 0{,}4 \times 0{,}5$$

   ambos realizan el mismo número de accesos, ya que el producto de las tasas de fallos de L2 y L3 coincide.

□

4. **(2 puntos)** Un procesador multinúcleo de 4.8GHz dispone de una memoria principal SDRAM DDR4 de un único DIMM de 64GB con buffers de fila de 8 KB. Es una DDR4-3200 con temporización 16-16-16 ($CL - t_{RCD} - t_{RP}$). El ancho del bus de memoria es de 64 bits.

El procesador ejecuta en el núcleo 0 la aplicación S de streaming que accede un vídeo ubicado en el mismo banco de memoria. La aplicación accede a todos los bloques de todas las filas de manera consecutiva una única vez.

El usuario ejecuta al mismo tiempo otra aplicación B en el núcleo 1 que accede al mismo banco pero a una fila distinta con una frecuencia de acceso menor. En concreto, una petición de la aplicación B se intercala cada 4 peticiones de la aplicación S en el controlador de memoria antes de ser lanzada al DIMM.

Se pide:

a) Indica la frecuencia del DIMM DDR4.

b) Tasa de acierto en el buffer de fila de la aplicación streaming S cuando se ejecuta sola para una cantidad de bloques accedida muy alta.

c) Penalización de fallos de la *cache* L3 en ciclos del procesador que experimenta la aplicación streaming S cuando se ejecuta sola asumiendo una tasa de aciertos en el buffer de fila del 100 %.

d) La aplicación S tarda más en ejecutarse cuando sus accesos a memoria se intercalan con los de la aplicación B. Razona cuál es el motivo.

e) Tasa de aciertos en el buffer de fila de la aplicación S cuando se ejecuta junto con la aplicación B.

Solución:

a) Indica la frecuencia del DIMM DDR4.

Como es una DDR4-3200, la frecuencia del módulo es de 1600 MHz.

b) Tasa de aciertos del buffer de fila cuando la aplicación streaming cuando se ejecuta sola.

$$\text{Numero de bloques buffer de fila} = \frac{8KB}{64B} = 128 \text{ bloques caben en el buffer.}$$

Si accede a todos los bloques de todas las filas, todos los accesos aciertan en el buffer de fila excepto el primer acceso.

$$\text{Tasa de aciertos buffer de fila} = \frac{127}{128} = 0{,}993$$

c) Penalización de fallos de la *cache* L3 en ciclos del procesador que sufre la aplicación streaming cuando se ejecuta sola asumiendo una tasa de aciertos en el buffer de fila del 100 %.

$$PF = \left(L \cdot (1 - TAbf) + L_r \cdot TAbf + \frac{B}{B_w} \right) \frac{f_{cpu}}{f_{mem}} \text{ ciclos de CPU}$$

Como TAbf = 1 entonces:

$$PF = \left(L_r \cdot TAbf + \frac{B}{B_w} \right) \frac{f_{cpu}}{f_{mem}} = \left(16 \cdot 1 + \frac{64}{16} \right) \frac{4800}{1600} = 60 \text{ ciclos de CPU}$$

d) La aplicación S tarda más en ejecutarse cuando sus accesos a memoria se intercalan con los de la aplicación B. Razona cuál es el motivo.

El motivo es que debe cerrarse la fila muchas más veces. En concreto, cada 4 accesos de A.

e) Tasa de aciertos del buffer de fila que experimenta la aplicación streaming cuando se ejecuta junto con la aplicación B.

Se intercala una petición de la aplicación B después de cada 4 peticiones de la aplicación A, por tanto, la secuencia de acceso es AAAABAAAAB...

Cada vez que llegue una petición de B debe cerrarse la fila para ubicar en el buffer la fila objetivo de la aplicación B. De las 4 peticiones de A, la primera falla abriendo la fila donde se encuentra parte del vídeo almacenado y las 3 restantes aciertan. En consecuencia, la tasa de aciertos para las peticiones de la aplicación A es del 75 %.

□