Ingeniería de Computadores
Enero 2013
Nombre: [redacted] **Grupo:**

Normas de realización:
- Incluir el nombre en todas las hojas utilizadas
- Todas las respuestas han de ser correctamente detalladas y razonadas.
- Las respuestas deben estar escritas con bolígrafo negro o azul

Un alumno de Ingeniería de Computadores del grupo ARA ha resuelto el examen que se presenta a continuación obteniendo únicamente 1 punto de los 10 posibles (es decir, solamente tiene una pregunta de 1 punto correcta). Resuelve correctamente las preguntas que ha contestado mal, indicando también cuál es la pregunta que está correcta.

Pregunta 1 (1 pto). Explica la utilidad del buffer de renombrado y enumera los tipos que hay según su direccionamiento.

Pregunta 2 (1 pto). Explica la diferencia entre predicción dinámica explícita y predicción dinámica implícita.

Pregunta 3 (1 pto). Justifica la diferencia que existe entre multicomputadores y multiprocesadores en términos de latencia y escalabilidad.

Pregunta 4 (1 pto). ¿Cuál es la característica distintiva de las redes de interconexión dinámicas frente a los otros tipos de redes de interconexión?

Pregunta 5 (1 pto). Indica a qué topología de la izquierda corresponden los conceptos de la derecha (N = número de nodos). Subraya las topologías que sean directas y a la vez irregulares.

A	Malla Illiac	C	Grado = 2
B	Barsil	E	$F_1(i) = (i - 1) \bmod r + (i \text{ DIV } r) \cdot r$, $N = r \cdot r$
C	Anillo	G	Ortogonal
D	Crossbar	F	$F_1(i) = (i - 1)$ si $i \bmod r <> 0$, $N = r \cdot r$
E	Toro	A	Diámetro = $q - 1$, $N = q \cdot q$
F	Malla abierta	H	Grado = 3
G	Hipercubo	B	Diámetro = $n / 2$, $n = \log N$
H	Árbol binario	D	Indirecta

Pregunta 6 (1 pto). Explica en qué consiste una arquitectura vectorial

Pregunta 7 (1 pto). Dibuja el diagrama de transiciones del protocolo de coherencia de caché MESI.

Pregunta 8 (1,5 ptos). Dada la siguiente secuencia de instrucciones:

```
lw r1, dato1 ; r1 = dato1
add r1,r1,r0 ; r1 = r1 + r0
lw r2, dato2 ; r2 = dato2
lw r3, dato3 ; r3 =dato3
add r4, r2, r3; r4= r2 + r3
mult r1, r1, r4; r1 = r1*r4
sub r2, r3, r1; r2 = r3 – r1
```

Indica cómo sería la evolución del buffer de reorden y de la tabla de registros. La arquitectura sobre la que se ejecutan las instrucciones está formada por las siguientes etapas: búsqueda de instrucción, decodificación/emisión, ejecución, reorden y escritura. El coste de cada etapa para todas las instrucciones es de 1 ciclo excepto en la etapa de ejecución. En ejecución hay dos unidades de carga/almacenamiento, una unidad de suma y una unidad de multiplicación; los tiempos de ejecución según el tipo de instrucción son los siguientes:

- Carga/almacenamiento: 1ciclo
- Suma/resta: 2 ciclos
- Multiplicación: 3 ciclos

La arquitectura puede captar y decodificar dos instrucciones en paralelo. También pueden finalizar dos instrucciones en paralelo. La estación de reserva es única para todas las unidades de ejecución y admite dos instrucciones como máximo y sigue una política de emisión no alineada.

Los valores iniciales de los registros son 0. Los datos son: datoA = 1, datoB = 4, datoC = 2

Pregunta 9 (1,5 ptos). El supercomputador UAS de la Universidad de Alicante consta de 1024 procesadores conectados en Anillo y es capaz de descomponer el 90% de cualquier aplicación para que sea ejecutada de forma paralela. A partir de los benchmarks del Servicio de Informática de la UA se constata que la paralelización que se consigue tiene las siguientes características:

- Un 50% de la parte paralelizable sólo puede utilizar 512 procesadores.
- Un 25% de la parte paralelizable sólo puede utilizar 256 procesadores.
- El resto de la parte paralelizable puede utilizar todos los procesadores.
- El coste de la comunicación es el 15% del tiempo de ejecución en paralelo en cada caso.

La Universidad de Alicante ha solicitado a IBM que estudie una actualización de su supercomputador para mejorar el rendimiento. Una propuesta de la compañía ha sido la de cambiar la red de comunicación considerando una conexión Hipercubo. Esta permite reducir el coste de la comunicación a un 2% del tiempo de ejecución en paralelo en cada caso. Sin embargo, la Universidad de Alicante no está dispuesta a realizar el desembolso económico propuesto con lo que ha pedido a IBM que estudie acelerar el 25% de la parte paralelizable cuando sólo puede utilizar 256 procesadores. Esto es, acelerando esa parte del proceso de forma equitativa para los 256 procesadores. ¿Cuántas veces más rápido tiene que ejecutarse, como mínimo, esa parte de las aplicaciones para que la solución sea competitiva con el cambio propuesto inicialmente por IBM?

RESOLUCIÓN

Pregunta 1

El buffer de renombrado es un buffer al que se accede desde la etapa de ejecución y que se utiliza para traducir los diferentes códigos de las instrucciones entre ensamblador y código máquina. Existen dos tipos de buffer de renombrado, el directo y el indirecto. Un buffer de renombrado directo es aquel que realiza una traducción entre ensamblador y código máquina sin

apoyo del compilador mientras que uno indirecto es el que utiliza el compilador para dicha traducción.
Un ejemplo de un buffer de renombrado sería la siguiente estructura (que incluye las traducciones de dos instrucciones ejemplo):

cmp r1, r2, r3	0x8956789
add r4, r5, r6	0xAC2394

Pregunta 2

La predicción dinámica explícita es aquella que se realiza en las primeras etapas del cauce, concretamente en la fase de prebúsqueda de la instrucción y, en ocasiones, también en la etapa de búsqueda de instrucción.
La predicción dinámica implícita es aquella que se realiza en la última etapa del cauce, antes de que la instrucción se complete.

Pregunta 3

En un sistema multicomputador, cada procesador dispone de su propio espacio de direcciones de memoria, accesible únicamente desde dicho procesador. Sin embargo, en un sistema multiprocesador, el espacio de direcciones es compartido entre los distintos procesadores. Esto tiene implicaciones en relación a la latencia y a la escalabilidad.

- Latencia en el acceso a memoria: es superior en los multiprocesadores, ya que el acceso concurrente a memoria compartida, a través de la red de interconexión, puede provocar aumentos en los tiempos de lectura y escritura en memoria.
- Escalabilidad con respecto al número de procesadores: es inferior en los multiprocesadores, ya que el rendimiento de los programas paralelos no aumentará en la misma proporción que el aumento de procesadores, debido a los conflictos de acceso a memoria.

Para paliar estas deficiencias presentes en los multiprocesadores frente a los multicomputadores, se han desarrollado arquitecturas multiprocesador que presentan no uniformidad en el acceso a memoria (NUMA) o memoria caché distribuida (COMA, CC-NUMA).

Pregunta 4

En una red de interconexión dinámica, los enlaces entre los nodos de la red son reconfigurables, esto es, es posible adaptar la topología de la red a la naturaleza de las aplicaciones en cada momento. Esto no ocurre en el resto de tipos de redes de interconexión, donde la topología se decide en tiempo de diseño y no puede cambiar.

Pregunta 5

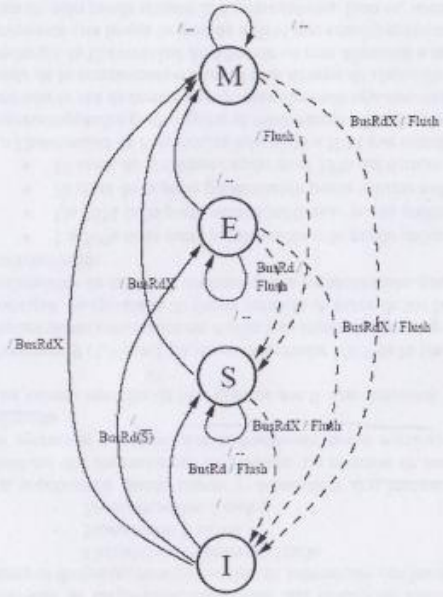
A	Malla Illiac	E	Grado = 2
B	Barril	B	$F_{-1}(i) = (i - 1) \bmod r + (i \text{ DIV } r) \cdot r$, $N = r \cdot r$
C	Anillo	F	Ortogonal
D	Crossbar	A	$F_{-1}(i) = (i - 1)$ si $i \bmod r <> 0$, $N = r \cdot r$
E	Toro	D	Diámetro = $q - 1$, $N = q \cdot q$
F	Malla abierta	G	Grado = 3
G	Hipercubo	C	Diámetro = $n / 2$, $n = \log N$
H	Árbol binario	H	Indirecta

Pregunta 6

Una arquitectura vectorial es una arquitectura secuencial orientada al procesamiento de vectores. El procesamiento de instrucciones está segmentado y se utilizan múltiples unidades funcionales superescalares. Esta se basa en el paralelismo funcional donde cada instrucción vectorial codifica una operación sobre todos los componentes del vector. Por tanto, dada operación vectorial codifica gran cantidad de cálculos permitiendo una reducción del número de instrucciones pero un aumento en el ciclo de reloj para la ejecución.

Pregunta 7

No recuerdo algunas transiciones…



Pregunta 8.

Ciclo 1

Las instrucciones #1, #2 y #3 (lw r1, dato1; add r1,r1,r0 ;lw r2, dato2) entran en el cauce en la etapa BI. Al entrar en BI entran en el buffer de reorden

Instr.	Cod. Op.	Reg. dest	Unidad Ex.	Result	Válido	Estado
#1	lw	R1	Carga1		0	Dentro
#2	add	R1	Suma1		0	Dentro
#3	lw	R2	Carga2		0	Dentro

Tabla de registros

Registro	Valor
R0	0
R1	0
R2	0
R3	0
R4	0

Ciclo 2

Las instrucciones #1, #2 y #3 pasan a la etapa de DI/ISS y la #4, #5 y #6 (lw r3, dato3 ; add r4, r2, r3; mult r1, r1, r4) entran en BI y por tanto en el buffer de reorden.

Instr.	Cod. Op.	Reg. dest	Unidad Ex.	Result	Válido	Estado
#1	lw	R1	Carga1		0	Dentro
#2	add	R1	Suma1		0	Dentro
#3	lw	R2	Carga2		0	Dentro
#4	lw	R3	Carga3		0	Dentro
#5	add	R4	Suma2		0	Dentro
#6	mult	R1	Mult1		0	Dentro

La tabla de registros no varia

Registro	Valor
R0	0
R1	0
R2	0
R3	0
R4	0

Ciclo 3

NO ME HA DADO TIEMPO A TERMINAR

Pregunta 9

Para poder comparar las propuestas de IBM, comenzamos calculando la ganancia con respecto al caso secuencial. Para ello, utilizaremos la Ley de Amdhal.

$$A = \frac{1}{1 - f_m + \frac{f_m}{A_m}} = \frac{1}{1 - 0.9 + \frac{0.9}{A_m}}$$

$$A_m = \frac{T_P}{\frac{0.5T_P}{512} + \frac{0.25T_P}{256} + \frac{0.25T_P}{1024} + 0.15\left(\frac{0.5T_P}{512} + \frac{0.25T_P}{256} + \frac{0.25T_P}{1024}\right)} =$$

$$A_m = \frac{T_P}{\frac{T_P + T_P + 0.25T_P}{1024} + 0.15\left(\frac{T_P + T_P + 0.25T_P}{1024}\right)} =$$

$$A_m = \frac{T_P}{\frac{2.25T_P}{1024} + 0.15\left(\frac{2.25T_P}{1024}\right)} = \frac{T_P}{\frac{2.25T_P}{1024} + \frac{0.3375T_P}{1024}} = \frac{T_P}{\frac{2.5865T_P}{1024}} =$$

$$A_m = 395.748792$$

$$A = \frac{1}{1 - f_m + \frac{f_m}{A_m}} = \frac{1}{1 - 0.9 + \frac{0.9}{395.748792}} = 9.77763986$$

El supercomputador UAS ejecuta las aplicaciones aproximadamente un 9.78 veces más rápido que las aplicaciones secuenciales. La propuesta inicial que tiene IBM es la de reducir el tiempo de comunicación utilizando una red Hipercubo. Esto aumentará la aceleración del sistema con respecto al caso secuencial. En este caso, en lugar de comparar con respecto al caso secuencial, calcularemos de nuevo la aceleración mejorada para compararla con la aceleración mejorada del supercomputador UAS.

$$A_{SOL1} = \frac{\frac{T_P + T_P + 0.25T_P}{1024} + 0.15\left(\frac{T_P + T_P + 0.25T_P}{1024}\right)}{\frac{T_P + T_P + 0.25T_P}{1024} + 0.02\left(\frac{T_P + T_P + 0.25T_P}{1024}\right)} =$$

$$A_{SOL1} = \frac{395.748792}{\frac{2.25T_P}{1024} + 0.02\left(\frac{2.25T_P}{1024}\right)} = \frac{395.748792}{\frac{2.25T_P}{1024} + \frac{0.045T_P}{1024}} = \frac{395.748792}{\frac{2.295T_P}{1024}} =$$

$$A_{SOL1} = \frac{395.748792}{0.00224121} = 176578.11$$

Ahora realizaremos el estudio que pide la UA a IBM de mejorar la parte del proceso que utiliza sólo los 256 procesadores. El cálculo lo realizaremos también con respecto a la aceleración sobre el supercomputador UAS.

$$A_{SOL2} = \frac{395.748792}{\frac{0.5T_P}{512} + \ldots} =$$

¡¡No me ha dado tiempo!! Para poder calcular cuántas veces más rápido tiene que ejecutarse como mínimo esa parte de las aplicaciones (para que pueda ser competitiva con el cambio propuesto inicialmente por IBM), la A_{SOL2} debería ser mayor o igual que la A_{SOL1}. Es decir:

$$A_{mSOL2} \geq A_{mSOL1}$$