

Arquitectura de Computadores



TEMA 3 Lanzamiento múltiple, Límites de ILP, Multithreading

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Curso 2015-2016

Contenidos

- o Introducción: CPI < 1
- o Lanzamiento múltiple de instrucciones: Superescalar, VLIW
- o Superescalar simple
- o VLIW
- o Superescalar con planificación dinámica
- o Límites de ILP
- o Ejemplo: Implementaciones X86
- o Thread Level Parallelism y Multithreading

- o Bibliografía

o Capítulo 3 y Anéndice H de [HePa12]

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

□ Introducción

- ¿Por que limitar a una instrucción por ciclo?
- Objetivo: $CPI < 1$
- Lanzar y ejecutar simultáneamente múltiples instrucciones por ciclo
- ¿Tenemos recursos?
 - Más área de silicio disponible
 - Técnicas para resolver las dependencias de datos (*planificación*)
 - Técnicas para resolver las dependencias de control (*especulación*)

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Más ILP: Lanzamiento múltiple

□ Alternativas

- Procesador Superescalar con planificación estática
- Procesador Superescalar con planificación dinámica+(especulación)
- Procesadores VLIW (very long instruction processor)

✓ Superescalar

- ✓ Lanza de 1 a 8 instrucciones por ciclo
- ✓ Reglas de ejecución
 - o Ejecución en orden-planificación estática
 - o Ejecución fuera de orden-planificación dinámica

✓ VLIW

- ✓ Numero fijo de instrucciones por ciclo

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Más ILP: Lanzamiento múltiple

□ Alternativas

Tipo	Forma del "issue"	Detección de riesgos	Planificación	Ejemplos
Superescalar estático	Dinámico	HW	estática	Embedded MIPS, ARM
Superescalar dinámico	Dinámico	HW	dinámica	ninguno
Superescalar especulativo	Dinámico	HW	Dinámica con especulación	P4, Core2, Power5, 7 SparcVI, VII
VLIW	Estático	Básicamente SW	estática	TI C6x Titanium

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Más ILP: Lanzamiento múltiple

□ SUPERESCALAR

Grado 2
2 Vías

Instruction	1	2	3	4	5	6	7
i	IF	ID	EX	MEM	WB		
i+1	IF	ID	EX	MEM	WB		
i+2		IF	ID	EX	MEM	WB	
i+3		IF	ID	EX	MEM	WB	
i+4			IF	ID	EX	MEM	WB
i+5			IF	ID	EX	MEM	WB

- **Duplicar todos los recursos:**
 - o Puertas bloque de Registros
 - o Fus
 - o Puertas de memoria,..
 - o Control
- **Ideal CPI= 0.5 se reduce por:**
 - o Saltos
 - o LOADs
 - o Dependencias verdaderas LDE
- **Necesita:**
 - o Predicción sofisticada

Efecto de riesgo LDE
provocado por un Load

Cartagena99

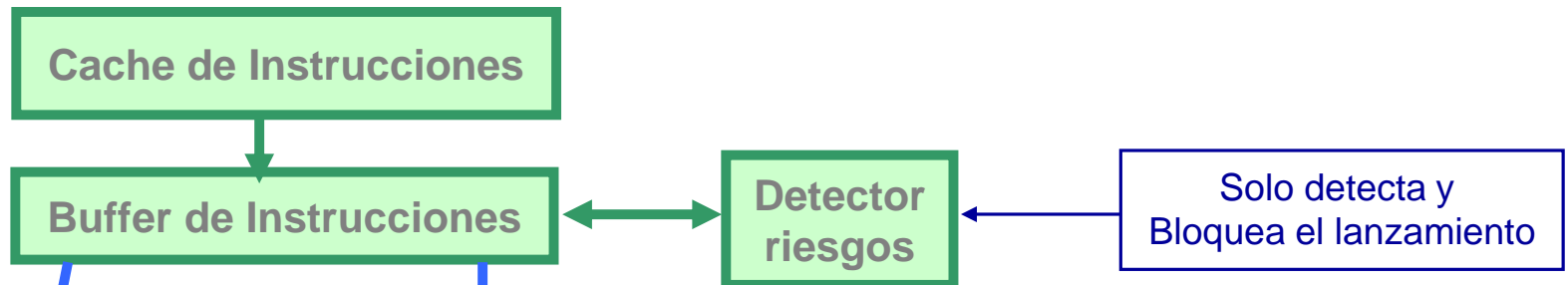
CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Más ILP: Lanzamiento múltiple

□ SUPERESCALAR Simple (estático, en orden)

- Regla de lanzamiento: Una instrucción FP (2ª)+ una instrucción de cualquier otro tipo (1ª)
- Buscar y decodificar dos instrucciones por ciclo (64 bits)
 - Ordenamiento y decodificación
 - Se analizan en orden. Sólo se lanza la 2ª si se ha lanzado la 1ª (conflictos)
- Unidades funcionales segmentadas (una ope. por ciclo) ó múltiples (división, raíz), más puertas en el bloque de registros
- Lanzamiento simple, recursos no conflictivos (diferentes reg y UF,..), excepto Conflictos de recursos; load, store, move FP → más puertas en el bloque de reg. Conflictos de datos LDE → más distancia entre instrucciones.



Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Más ILP: Lanzamiento múltiple

□ SUPERESCALAR Simple (estático, en orden)

	<u>Instrucción entera</u>	<u>Instrucción FP</u>	<u>Ciclo</u>
Loop:	LD F0,0(R1)		1
	LD F6,-8(R1)		2
	LD F10,-16(R1)	ADDD F4,F0,F2	3
	LD F14,-24(R1)	ADDD F8,F6,F2	4
	LD F18,-32(R1)	ADDD F12,F10,F2	5
	SD 0(R1),F4	ADDD F16,F14,F2	6
	SD -8(R1),F8	ADDD F20,F18,F2	7
	SD -16(R1),F12		8
	SUBI R1,R1,#40		9
	SD 16(R1),F16		10
	BNEZ R1,LOOP		11
	SD 8(R1),F20		12

Separadas por 2 ciclos

- Desarrollo para ejecución superescalar: se desarrolla una iteración más

CLASES PARTICULARES, TUTORIAS TECNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Cartagena99

Más ILP: Lanzamiento múltiple

□ SUPERESCALAR Simple (estático, en orden)

➤ *Ventajas*

- No modifica código. Compatibilidad binaria
- No riesgos en ejecución

➤ *Desventajas*

- Mezcla de instrucciones. Solo obtiene CPI de 0.5 en programas con 50 % de FP
- Bloqueos en el lanzamiento
- Planificación fija: No puede adaptarse a cambios en ejecución (Fallos de cache)
- Los códigos deben de ser replanificados para cada nueva implementación (eficiencia)

Cartagena99

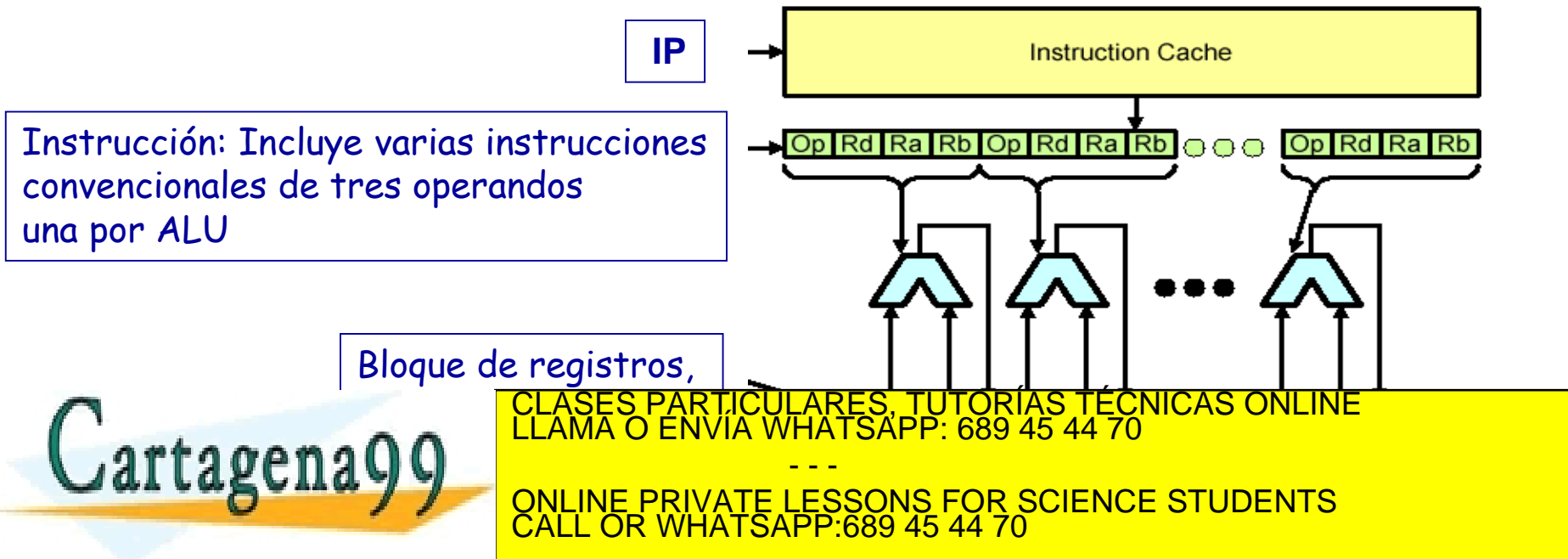
CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Más ILP: Lanzamiento múltiple

□ VLIW

- El análisis de dependencias en tiempo de compilación
- Muchas operaciones por instrucción (IA64 packet, Transmeta molécula)
- Todas las operaciones de una instrucción se ejecutan en paralelo
- Instrucciones con muchos bits
- Muchas operaciones vacías (NOP)



Cartagena99

CLASES PARTICULARES, TUTORIAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Más ILP: Lanzamiento múltiple

□ VLIW Ejemplo Tema 2

```
LOOP   LD      F0,0(R1)
        ADDD   F4,F0,F2
        SD     0(R1),F4
        SUBI   R1,R1,#8
        BNEZ   R1,LOOP
```

- Aplicar técnicas conocidas para minimizar paradas
 - Unrolling
 - Renombrado de registros

- Latencias de uso: LD a ADD 1 ciclo, ADD a SD 2 ciclos

```
LOOP:  LD      F0, 0(R1)
        LD      F6, -8(R1)
        LD      F10, -16(R1)
        LD      F14, -24(R1)
        ADDD   F4, F0, F2
        ADDD   F8, F6, F2
        ADDD   F12, F10, F2
        ADDD   F16, F14, F2
        SD     0(R1), F4
        SD     -8(R1), F8
        SUBI   R1, R1, #32
        SD     -16(R1), F12
```



CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Cartagena99

Más ILP: Lanzamiento múltiple

□ VLIW

Loop unrolling en VLIW (desarrollo 7 iteraciones)

```

LOOP:  LD F0,0(R1)           ; F0 = array element
        ADDD F4,F0,F2        ; add scalar in F2
        SD 0(R1),F4         ; store result
        SUBI R1,R1,#8       ; decrement pointer
        BNEZ R1, LOOP       ; branch if R1!=0
    
```

<u>Mem ref 1</u>	<u>Mem ref 2</u>	<u>FP op</u>	<u>FP op</u>	<u>Int op/branch</u>
LD F0,0(R1)	LD F6,-8(R1)			
LD F10,-16(R1)	LD F14,-24(R1)			
LD F18,-32(R1)	LD F22,-40(R1)	ADDD F4,F0,F2	ADDD F8,F6,F2	
LD F26,-48(R1)		ADDD F12,F10,F2	ADDD F16,F14,F2	
		ADDD F20,F18,F2	ADDD F24,F22,F2	
SD 0(R1),F4	SD -8(R1),F8	ADDD F28,F26,F2		
SD -16(R1),F12	SD -24(R1),F16			SUBI R1,R1,#56
SD 24(R1),F20	SD 16(R1),F24			

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
 LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
 CALL OR WHATSAPP:689 45 44 70

... registros necesarios

Más ILP: Lanzamiento múltiple

□ VLIW

VENTAJAS

- Hardware de control muy simple
 - No detecta dependencias
 - Lógica de lanzamiento simple
- Puede explotar paralelismo a todo lo largo del programa (compilador)

DESVENTAJAS

- Planificación estática; Muy sensible a fallos de cache
- Necesita desenrollado muy agresivo
- Bloque de registros muy complejo en área y tiempo de acceso
- Muchas NOP
 - Poca densidad de código
 - Capacidad y AB de la cache de instrucciones

Cartagena99

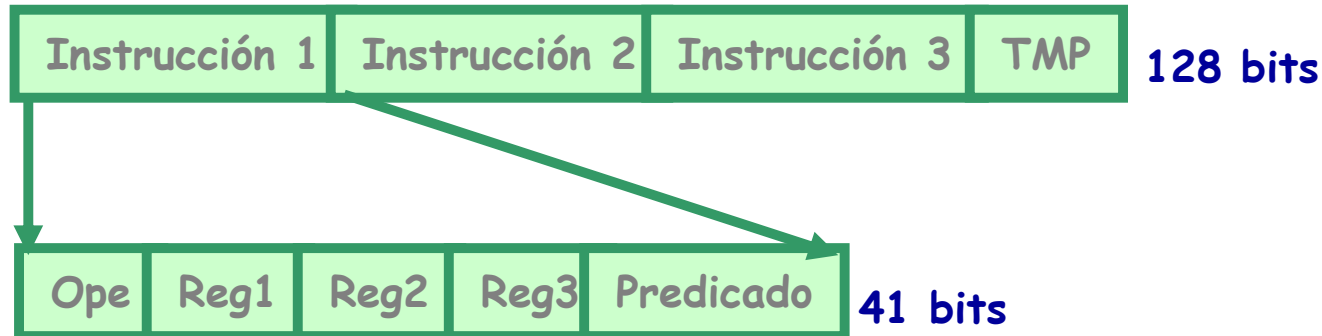
CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Más ILP: Lanzamiento múltiple

□ EPIC: Explicitly Parallel Instruction Computing (IA64)

- Instrucciones de 128 bits
 - Operaciones de tres operandos
 - TMP (Template) codifica dependencias entre las operaciones
 - 128 registros enteros (64bits), 128 registros FP (82bits)
 - Ejecución predicada. 64 registros de predicado de 1 bit
 - Cargas especulativas
 - Hw para chequeo de dependencias



Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Más ILP: Lanzamiento múltiple

□ SUPERESCALAR con Planificación Dinámica. Fuera de orden

➤ Un Diseño Simple

- Estaciones de reserva separadas para enteros (+reg) y PF (+reg)
- Lanzar dos instrucciones en orden (ciclo de lanzamiento: partir en dos subciclos)
- Solo FP load causan dependencias entre instrucciones enteras y PF
 - Reemplazar buffer de load con cola. Las lecturas se hacen en orden
 - Ejecución Load: "check" dirección en cola de escritura para evitar LDE
 - Ejecución Store: "check" dirección en cola de lecturas para evitar EDL

➤ Rendimiento del procesador

<u>Iteración no.</u>	<u>Instrucción</u>	<u>Lanzada</u>	<u>Comienza Ejecución</u> (número de ciclo)	<u>Escribe resultado</u>
1	LD F0,0(R1)	1	2	4
1	ADDD F4,F0,F2	1	5	8 ←
1	SD 0(R1),F4	2	9	
1	SUBI R1,R1,#8	3	4	5
1	BNEZ R1,LOOP	4	6	
2	LD F0,0(R1)	5	6	8 ←

4 ciclos por iteración

Cartagena99

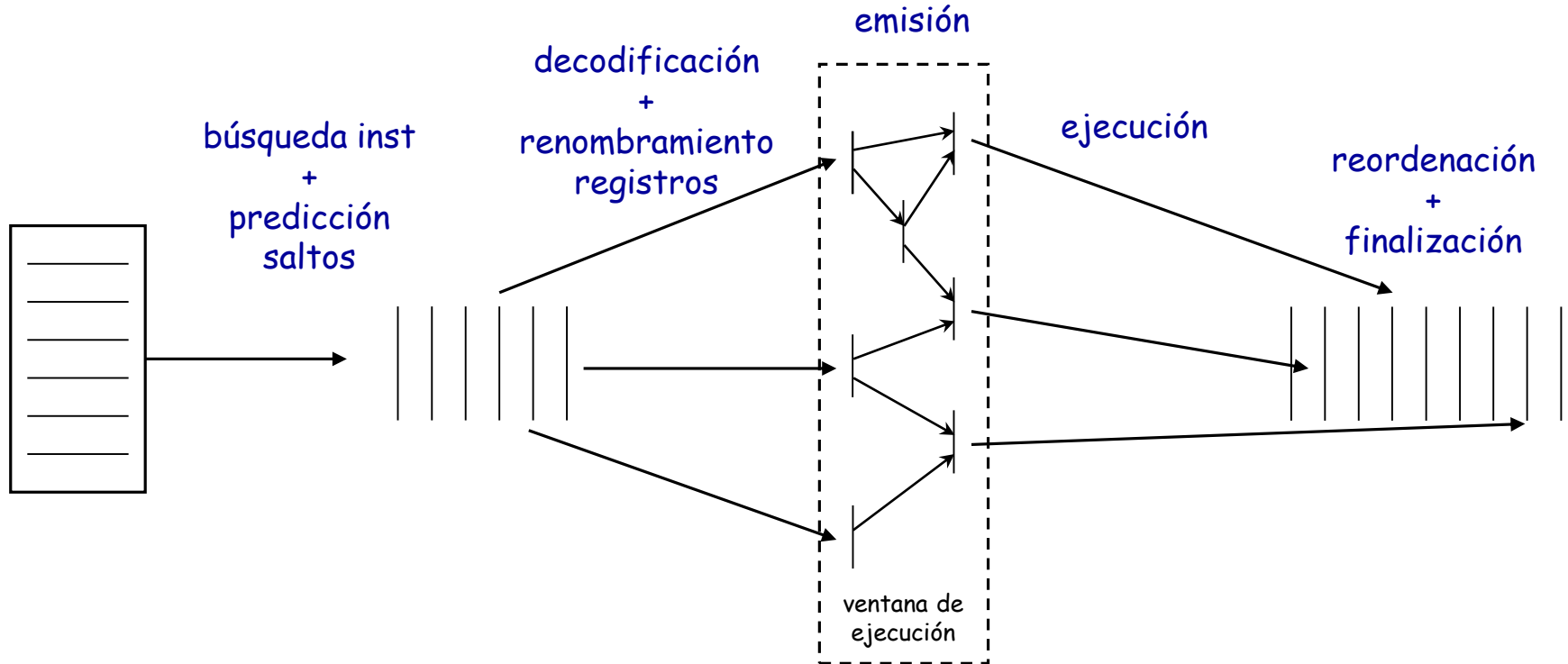
CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Más ILP: Lanzamiento múltiple

□ SUPERESCALAR con Planificación Dinámica y Especulación

Ejecución fuera de orden. Finalización en orden



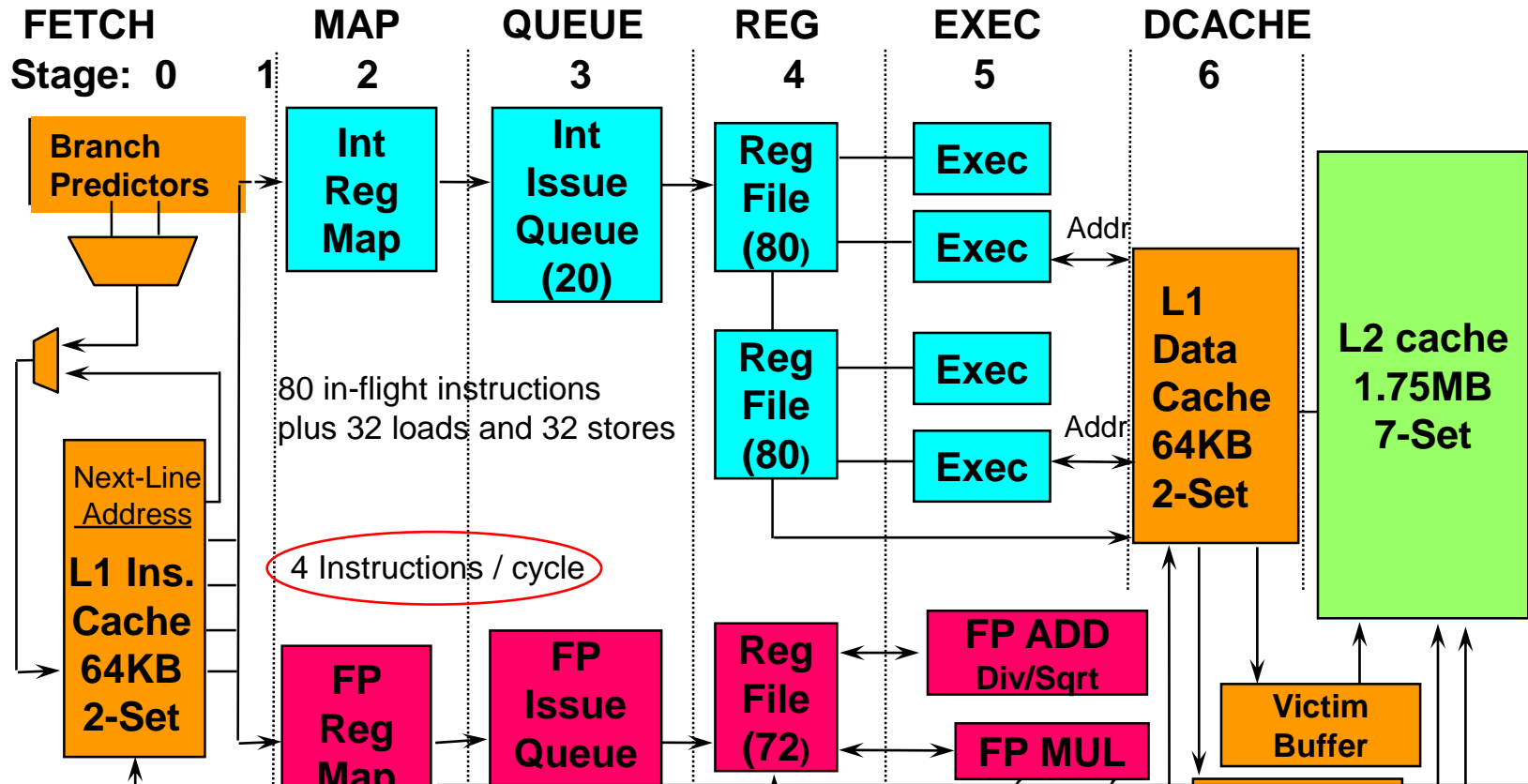
Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP: 689 45 44 70

Más ILP: Lanzamiento múltiple

□ EV7 ALPHA 21364 Core (2003)



Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP: 689 45 44 70

Límites del ILP

- ❑ El lanzamiento múltiple permite mejorar el rendimiento sin afectar al modelo de programación
- ❑ En los últimos años se ha mantenido el mismo ancho superescalar que tenían los diseños del 1995
- ❑ La diferencia entre rendimiento pico y rendimiento obtenido crece
- ❑ ¿Cuanto ILP hay en las aplicaciones?
- ❑ ¿Necesitamos nuevos mecanismos HW/SW para explotarlo?
 - o Extensiones multimedia:

o Total MMY SSE SSE2 SSE3 SSE4

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Límites del ILP

- ❑ ¿Cuanto ILP hay en las aplicaciones?

- ❑ Supongamos un procesador superescalar fuera de orden con especulación y con recursos ilimitados
 - o Infinitos registros para renombrado
 - o Predicción perfecta de saltos
 - o Caches perfectas
 - o Lanzamiento no limitado
 - o Desambiguación de memoria perfecta

- ❑ Entonces...
 - o En tal sistema la ejecución de instrucciones está solo

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Límites del ILP

- Comparaciones necesarias en lanzamiento
 - o Ejemplo: Lanzar 6 instrucciones/ciclo (check LDE)
 - 1-2 1-3 1-4 1-5 1-6 = 5 instr x 2 operandos = 10 comp
 - 2-3 2-4 2-5 2-6 = 4 instr x 2 operandos = 8 comp
 -
 - 5-6 = 1 instr x 2 operandos = 2 comp
 - TOTAL = 30 comp
 - o En general: lanzar n instrucciones → n^2-n comp

- Comparaciones necesarias para paso de operandos
 - o Ejemplo: pueden acabar 6 instr/ciclo, 80 instr en la ventana, 2 operandos/instr.
 - $6 \times 80 \times 2 = 960$ comparaciones/ciclo
 - o En general, comparar con los operandos de todas las

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP: 689 45 44 70

Límites del ILP

❑ Modelo versus procesador real

	Modelo	Power 5
Instrucciones lanzadas por ciclo	Infinito	4
Ventana de instrucciones	Infinita	200
Registros para renombrado	Infinitos	48 integer + 40 Fl. Pt.
Predicción de saltos	Perfecta	2% to 6% de fallos de predicción (Tournament Branch Predictor)
Cache	Perfecta	64KI, 32KD, 1.92MB L2, 36 MB L3 (off-chip)

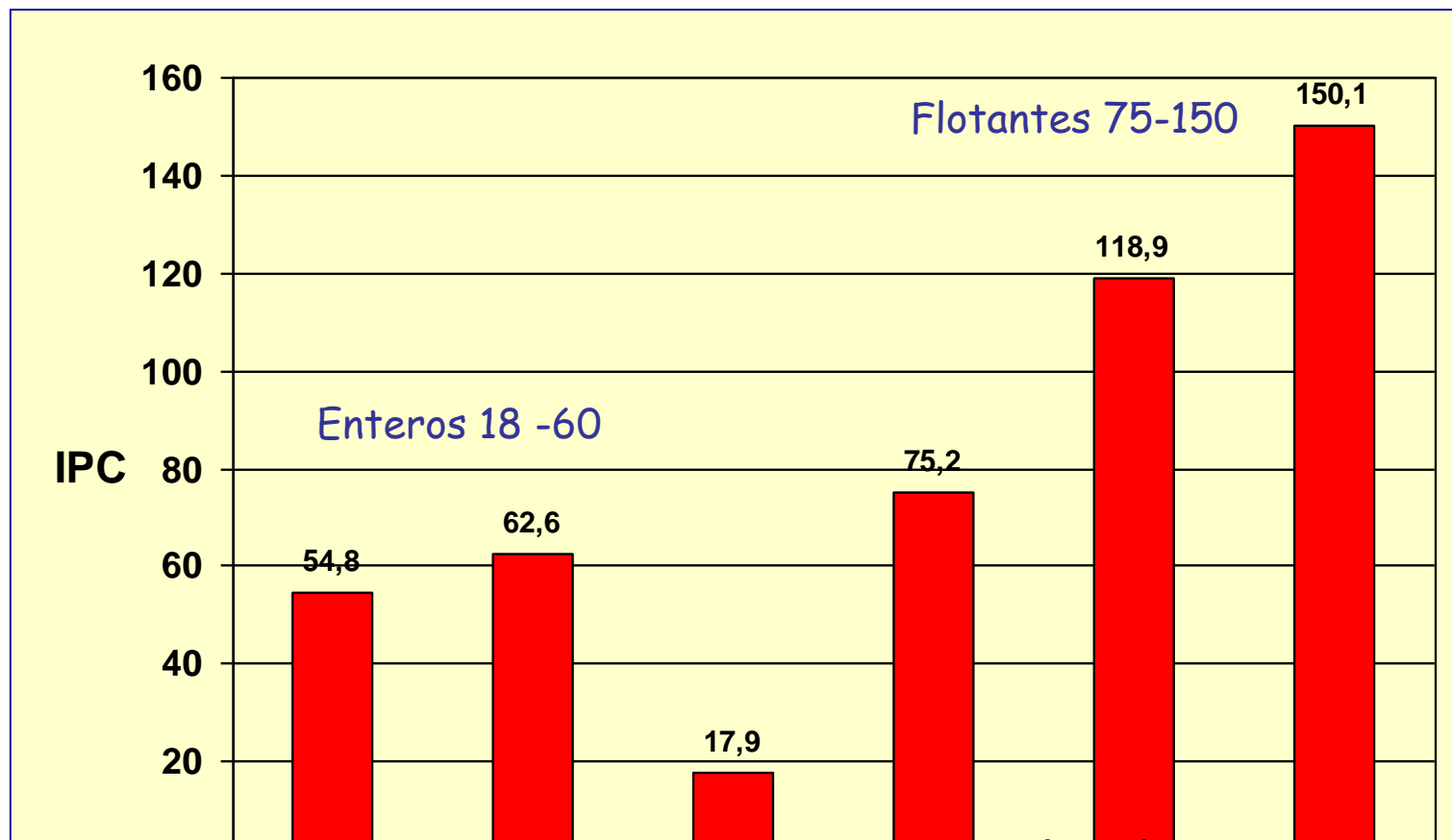
Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Límites del ILP

□ Límite superior: Recursos ilimitados



Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Límites del ILP

❑ Modelo versus procesador real

	Nuevo Modelo	Modelo	Power 5
Instrucciones lanzadas por ciclo	Infinitas	Infinitas	4
<u>Ventana de instrucciones</u>	Infinito vs. 2K, 512, 128, 32, 8, 4	Infinita	200
Registros para renombrado	Infinitos	Infinitos	48 enteros + 40 Fl. Pt.
Predicción de saltos	Perfecta	Perfecta	Tournament
Cache	Perfecta	Perfecta	64KI, 32KD, 1.92MB L2, 36 MB L3

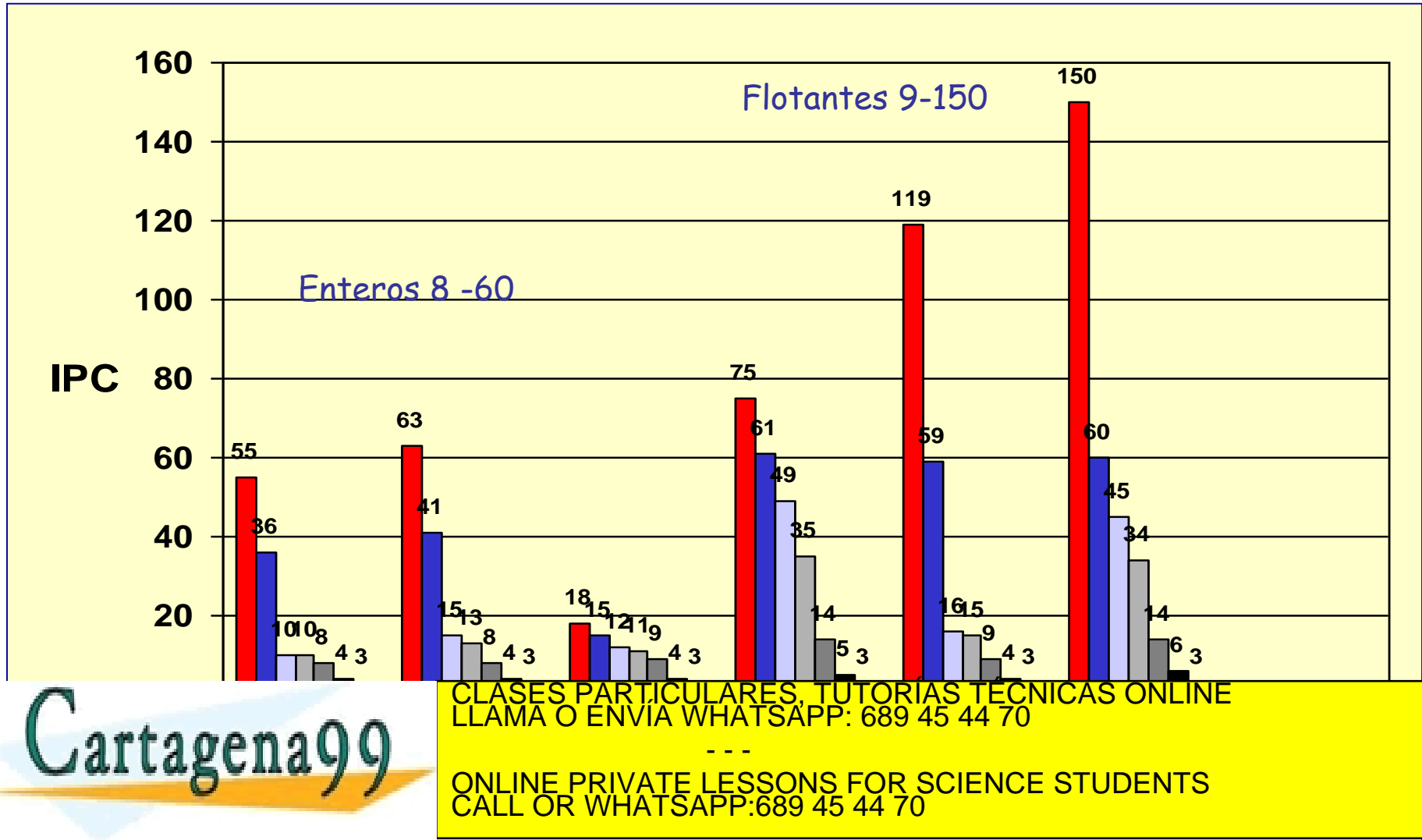
Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP: 689 45 44 70

Límites del ILP

□ HW más real: Impacto del tamaño de la ventana de instrucciones



Cartagena99

CLASES PARTICULARES, TUTORIAS TECNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Límites del ILP

❑ Modelo versus procesador real

	Nuevo Modelo	Modelo	Power 5
Instrucciones lanzadas por ciclo	64 sin restricciones	Infinitas	4
Ventana de instrucciones	2048	Infinita	200
Registros para renombrado	Infinitos	Infinitos	48 enteros + 40 Fl. Pt.
<u>Predicción de saltos</u>	Perfecto vs. 8K Tournament vs. 512 2-bit vs. profile vs. ninguno	Perfecta	Tournament
Cache	Defecto	Defecto	64KT 32KD 102MD

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
 LLAMA O ENVÍA WHATSAPP: 689 45 44 70

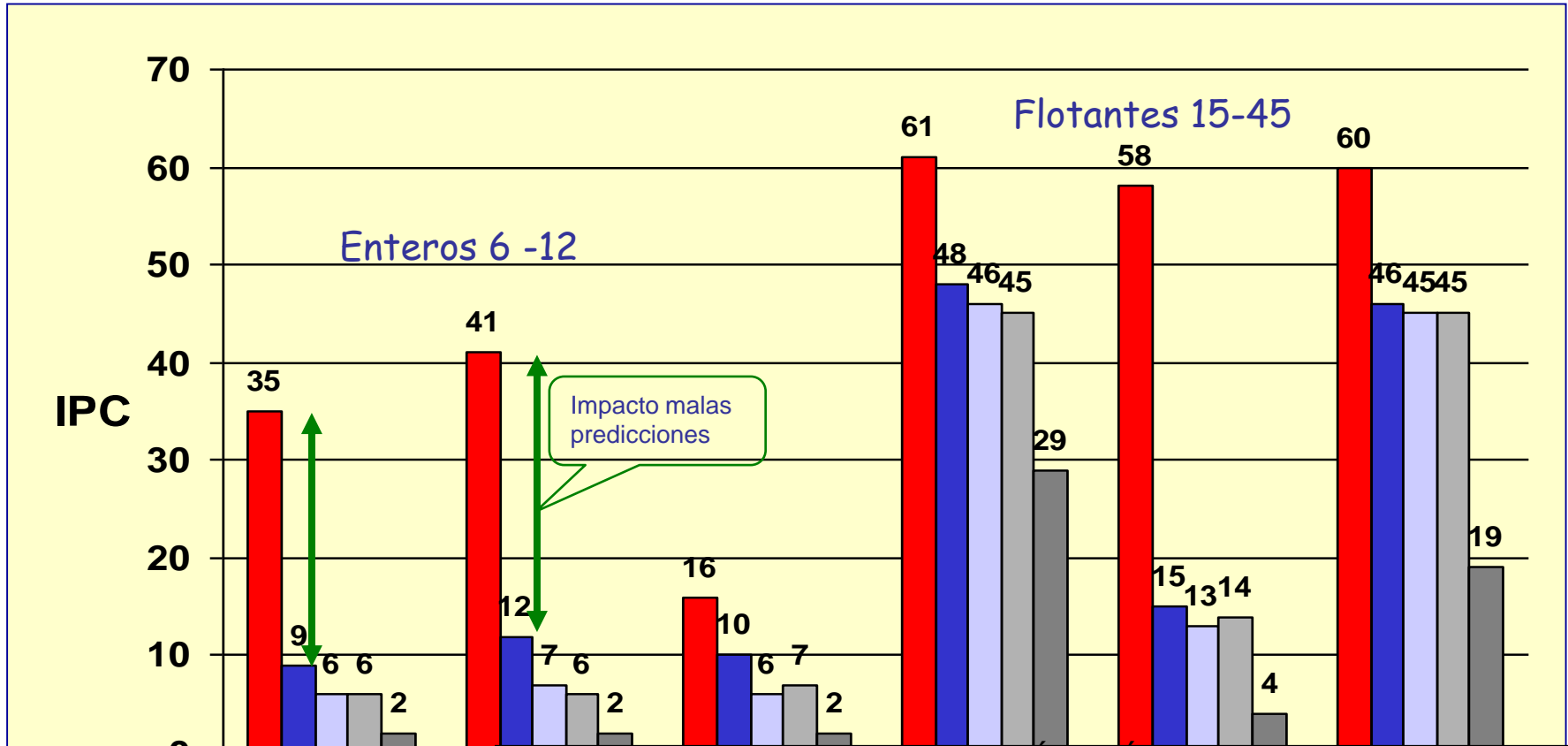
ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
 CALL OR WHATSAPP:689 45 44 70

Cartagena99

Límites del ILP

☐ HW más real: Impacto de los saltos

Ventana de 2048 y lanza 64 por ciclo



CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Cartagena99

Límites del ILP

❑ Modelo versus procesador real

	Nuevo Modelo	Modelo	Power 5
Instrucciones lanzadas por ciclo	64 sin restricciones	Infinitas	4
Ventana de instrucciones	2048	Infinita	200
<u>Registros para renombrado</u>	Infinito v. 256, 128, 64, 32, ninguno	Infinitos	48 enteros + 40 Fl. Pt.
Predicción de saltos	8K Tournament (hibrido)	Perfecta	Tournament
Cache	Perfecta	Perfecta	64KI, 32KD, 1.92MB L2, 36 MB L3

Cartagena99

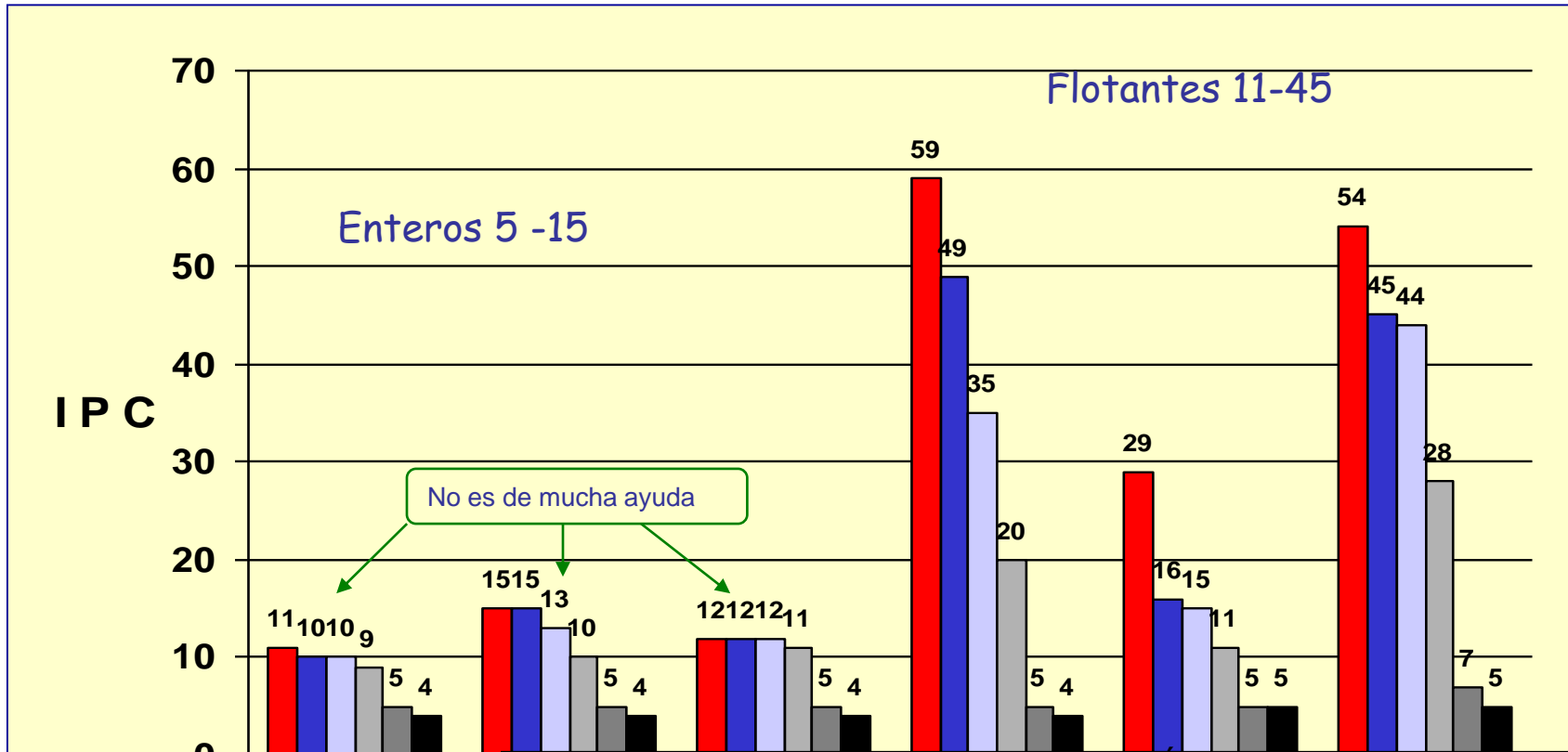
CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Límites del ILP

□ HW más real: Impacto del numero de registros

Ventana de 2048 y lanza 64 por ciclo, predictor híbrido 8K entradas



No es de mucha ayuda

CLASES PARTICULARES, TUTORIAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP: 689 45 44 70

Cartagena99

Límites del ILP

❑ Modelo versus procesador real

	Nuevo Modelo	Modelo	Power 5
Instrucciones lanzadas por ciclo	64 (sin restricciones)	Infinitas	4
Ventana de instrucciones	Infinito vs. 256, 128, 32, 16	Infinita	200
Registros para renombrado	64 Int + 64 FP	Infinitos	48 enteros + 40 Fl. Pt.
Predicción de saltos	1K 2-bit	Perfecta	Tournament
Cache	Perfecto	Perfecta	64KI, 32KD, 1.92MB L2, 36 MB L3
Análisis de	LHW/	Perfecto	

Cartagena99

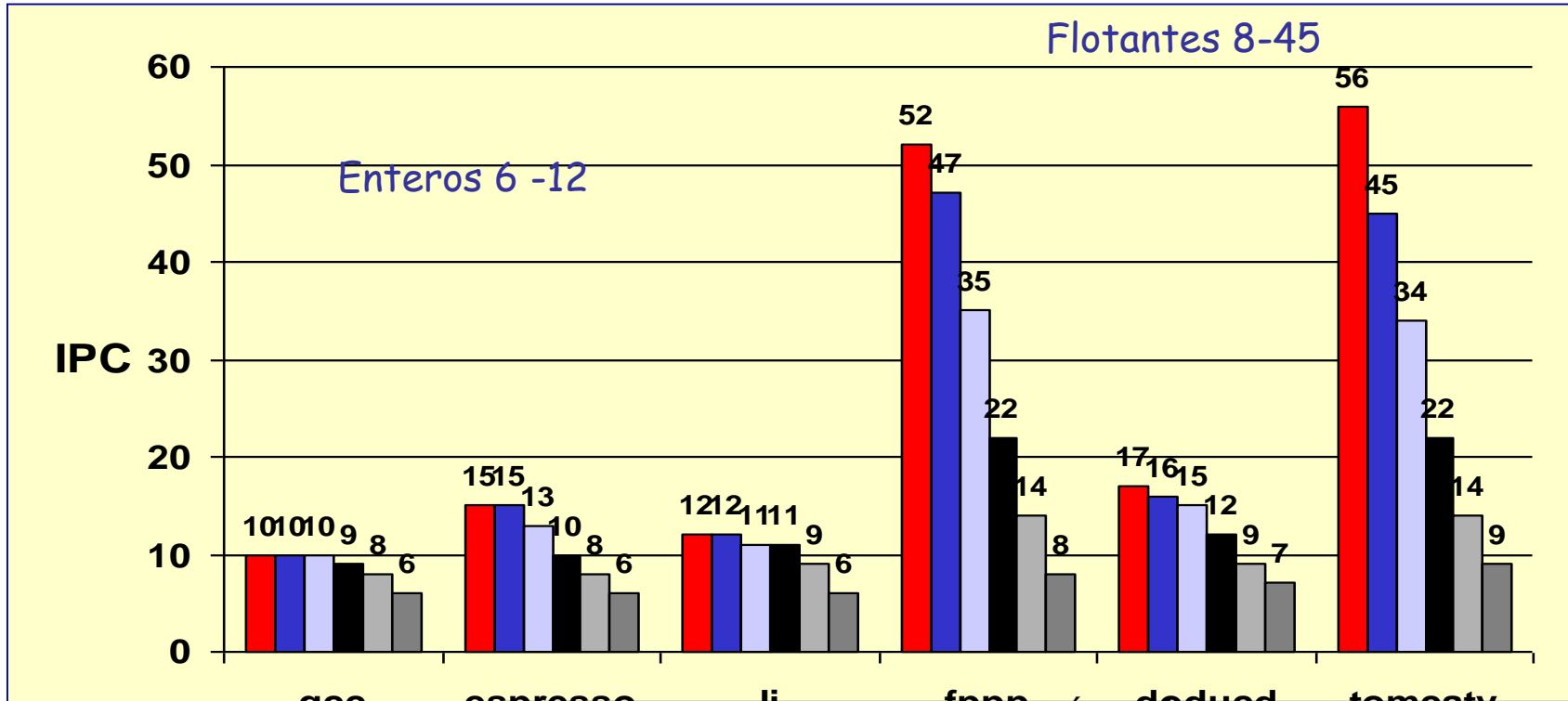
CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Límites del ILP

☐ HW realizable

64 instrucciones por ciclo sin restricciones, predictor híbrido, predictor de retornos de 16 entradas, Desambiguación perfecta, 64 registros enteros y 64 Fp extra



Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Un ejemplo

Implementaciones X86: P6, Netburst(Pentium4)...

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

- P6 (implementacion de la arquitectura IA-32 usada en Pentium Pro, II, III)

Modelo	Año	Clock	L1	L2
Pentium Pro	1995	100-200Mhz	8+8 KB	126-1024KB
Pentium II	1998	233-450Mhz	16+16 KB	256-512KB
Celeron	1999	500-900Mhz	16+16 KB	128KB
Pentium III	1999	450-1100Mhz	16+16 KB	256-512KB
PentiumIII Xeon	1999	700-900Mhz	16+16 KB	1MB-2MB

¿ Como segmentar un ISA con instrucciones entre 1 y 17 bytes?

- o El P6 traduce las instrucciones originales IA-32 a "microoperaciones" de 118 bits (Traducción Dinámica)
 - Formato μ -op: cod+3 operandos. Modelo: load/store
 - Muy descodificadas
- o Cada instrucción original se traduce a una secuencia de 1 a 4 microoperaciones.

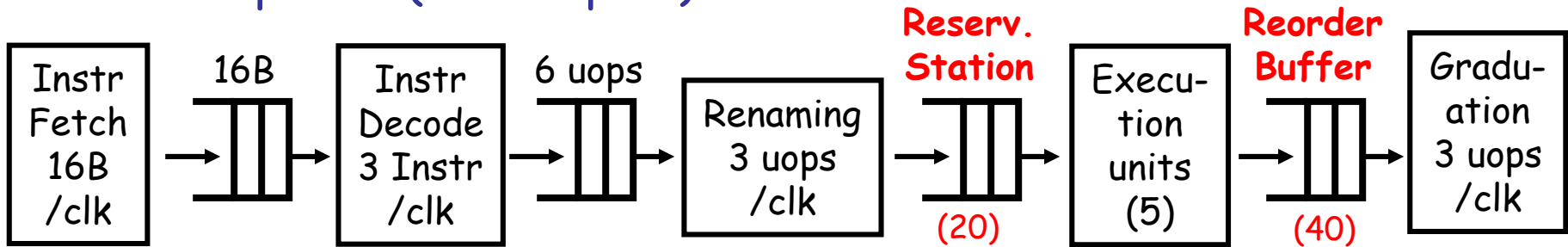
Pern

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

□ P6 Pipeline (14 etapas)



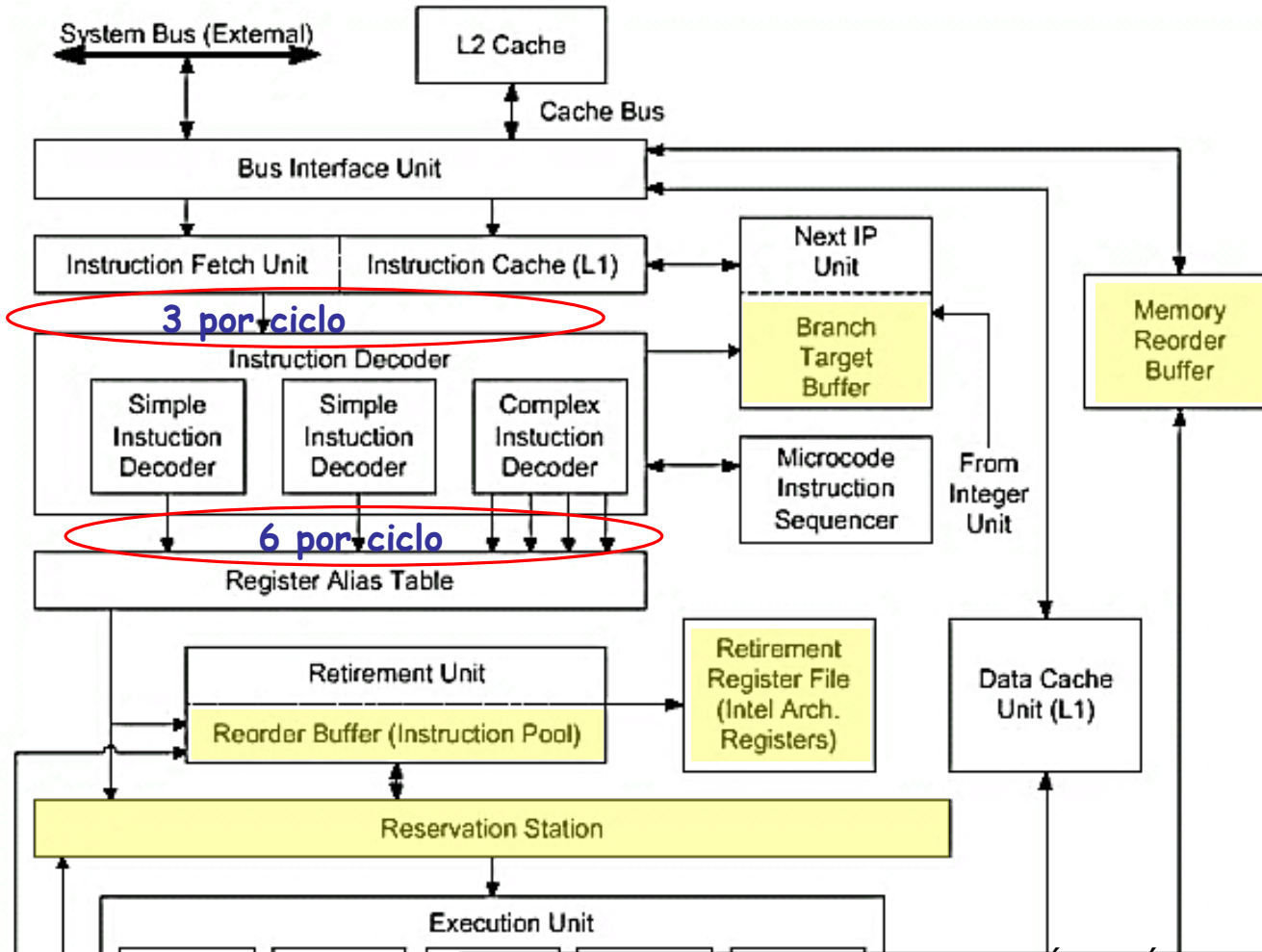
- 8 etapas para fetch, decodificación y issue en orden
 - 1 ciclo para determinar la longitud de la instrucción 80x86 + 2 más para generar las microoperaciones
- 3 etapas para ejecución fuera de orden en una de 5 unidades funcionales
- 3 etapas para la finalización de la instrucción (commit)

<u>Parameter</u>	<u>80x86</u>	<u>microops</u>
Max. instructions issued/clock	3	6
Max. instr. complete exec./clock		5
Max. instr. committed/clock		3
Window (Instrs in reorder buffer)	40	
Number of reservations stations	20	
Number of reserved registers	40	

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
 LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
 CALL OR WHATSAPP:689 45 44 70



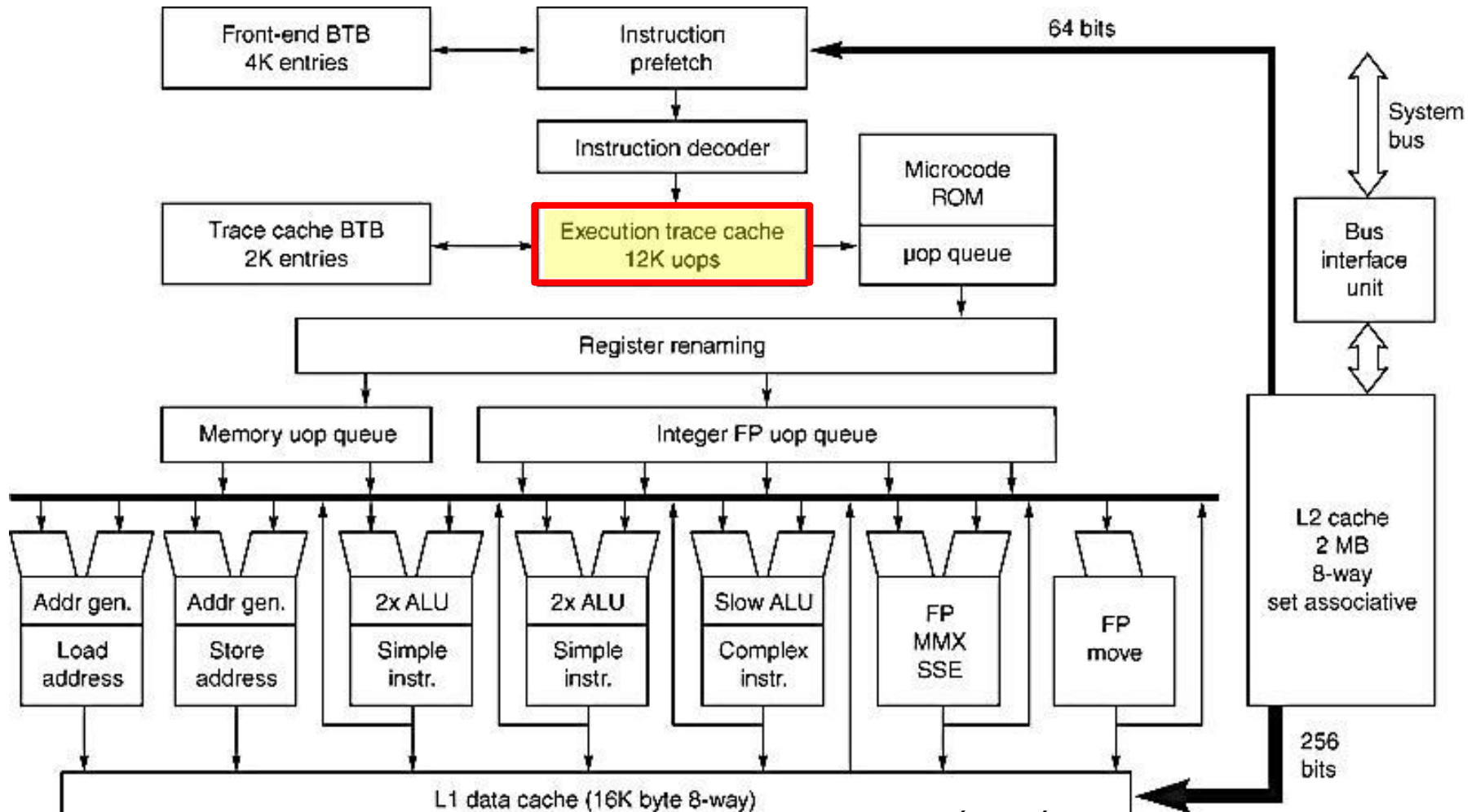
CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
 LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
 CALL OR WHATSAPP: 689 45 44 70

Internal Data-Results Buses

Cartagena99

Pentium 4 (Netburst Microarchitecture)



CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
 LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
 CALL OR WHATSAPP:689 45 44 70

Cartagena99

☐ "Double pumped ALU" means ALU clock rate 2X => 2X ALU F.U.s

Pentium 4

Intel Netburst Microarchitecture

- ❑ Traduce instrucciones 80x86 a micro-ops (como P6)
- ❑ P4 tiene mejor predicción de saltos (x8) y más FU (7 versus 5)
- ❑ La Cache de instrucciones almacena micro-operaciones vs. 80x86 instrucciones.
"trace cache" (TC), BTB TC 2K entradas
 - o En caso de acierto elimina decodificación
- ❑ Nuevo bus de memoria: 400(800) MHz vs. 133 MHz (RamBus, DDR, SDRAM)
(Bus@1066 Mhz)
- ❑ Caches
 - o Pentium III: L1I 16KB, L1D 16KB, L2 256 KB
 - o Pentium 4: L1I 12K uops, L1D 16 KB 8-way, L2 2MB 8-way
- ❑ Clock :
 - o Pentium III 1 GHz v. Pentium 4 1.5 GHz (3.8 Ghz)
 - o 14 etapas en pipeline vs. 24 etapas en pipeline (31 etapas)
- ❑ Instrucciones Multimedia: 128 bits vs. 64 bits => 144 instrucciones nuevas.
- ❑ Usa RAMBUS DRAM

Cartagena99

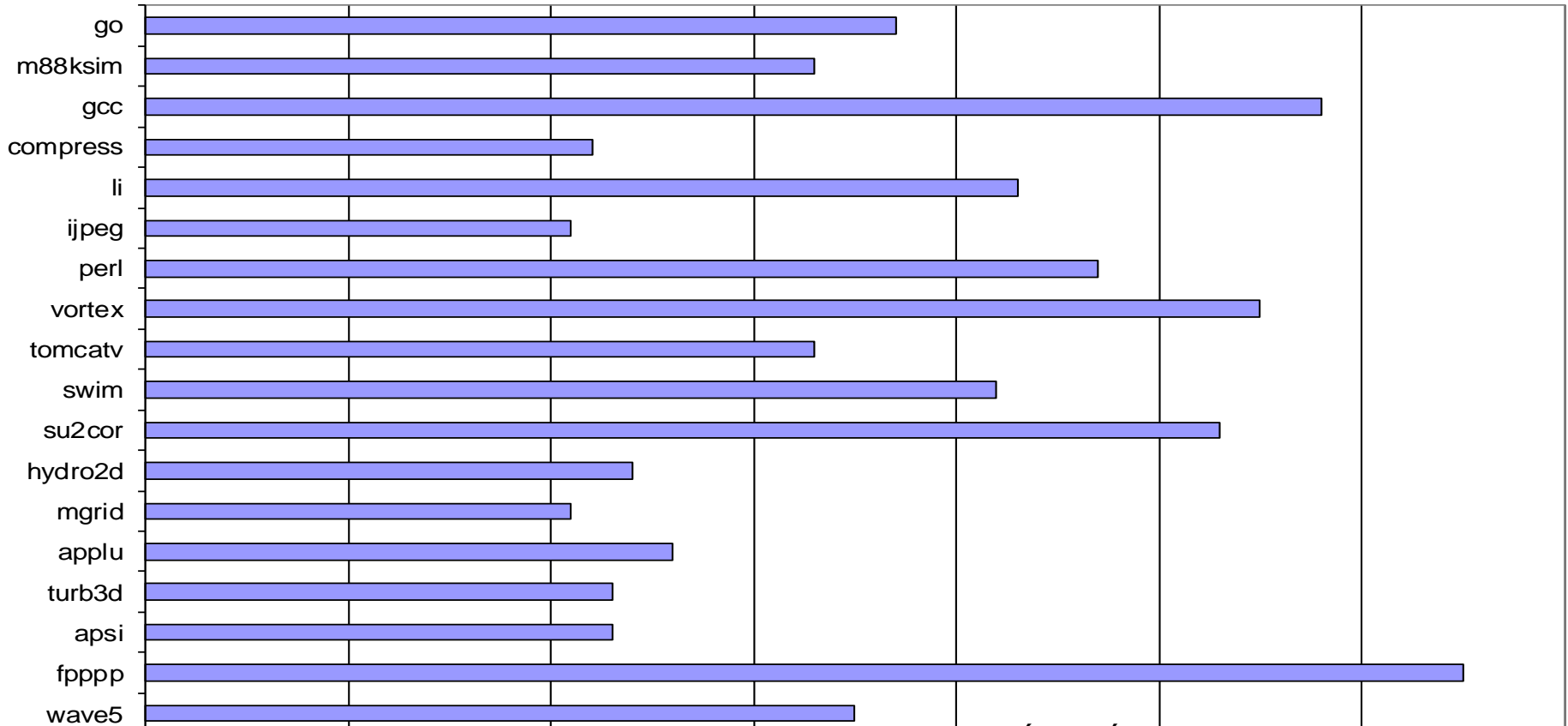
CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

❑ DID: 512 vs. 1024 entradas. Mejora 50% la tasa de malas predicciones

Netburst(Pentium 4) Rendimiento

Relación μ -operaciones / instrucciones x86



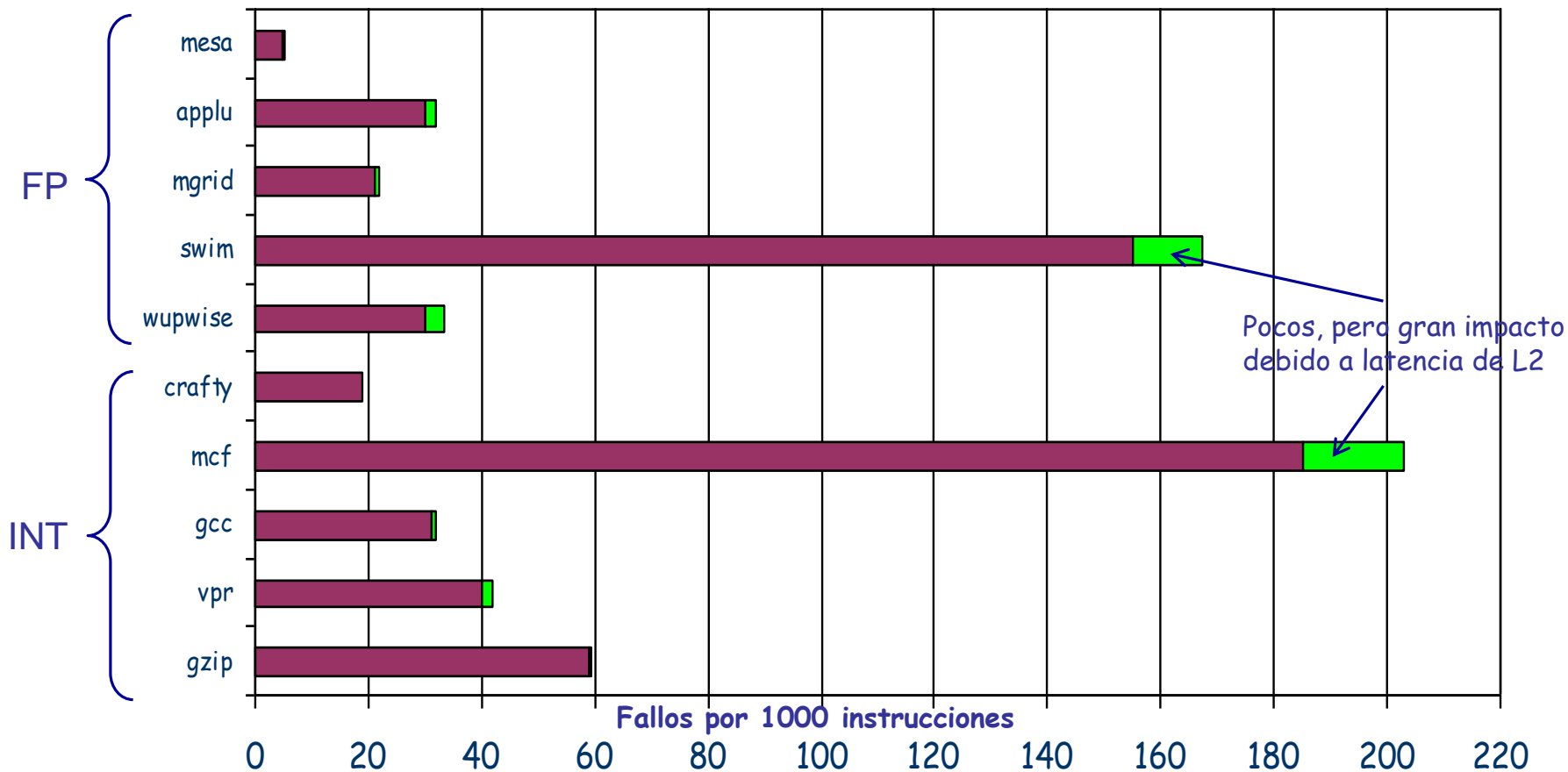
Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Netburst(Pentium4) Rendimiento

Fallos de cache de datos



CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

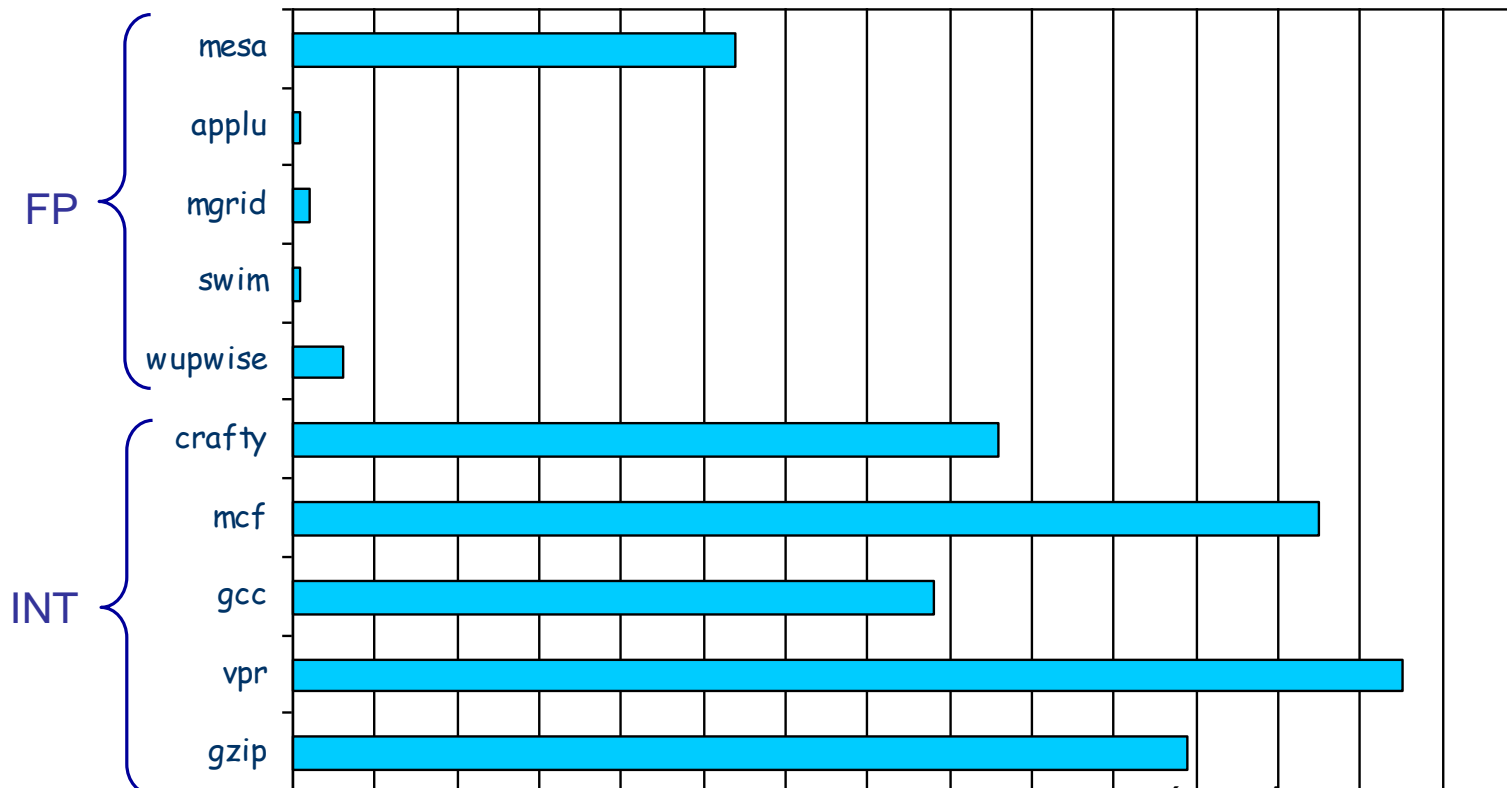
ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

De 10 a 200 fallos por cada mil instrucciones



Netburst(Pentium4) Rendimiento

Comportamiento del predictor: Dos niveles, con información global y local
4K entradas



Cartagena99

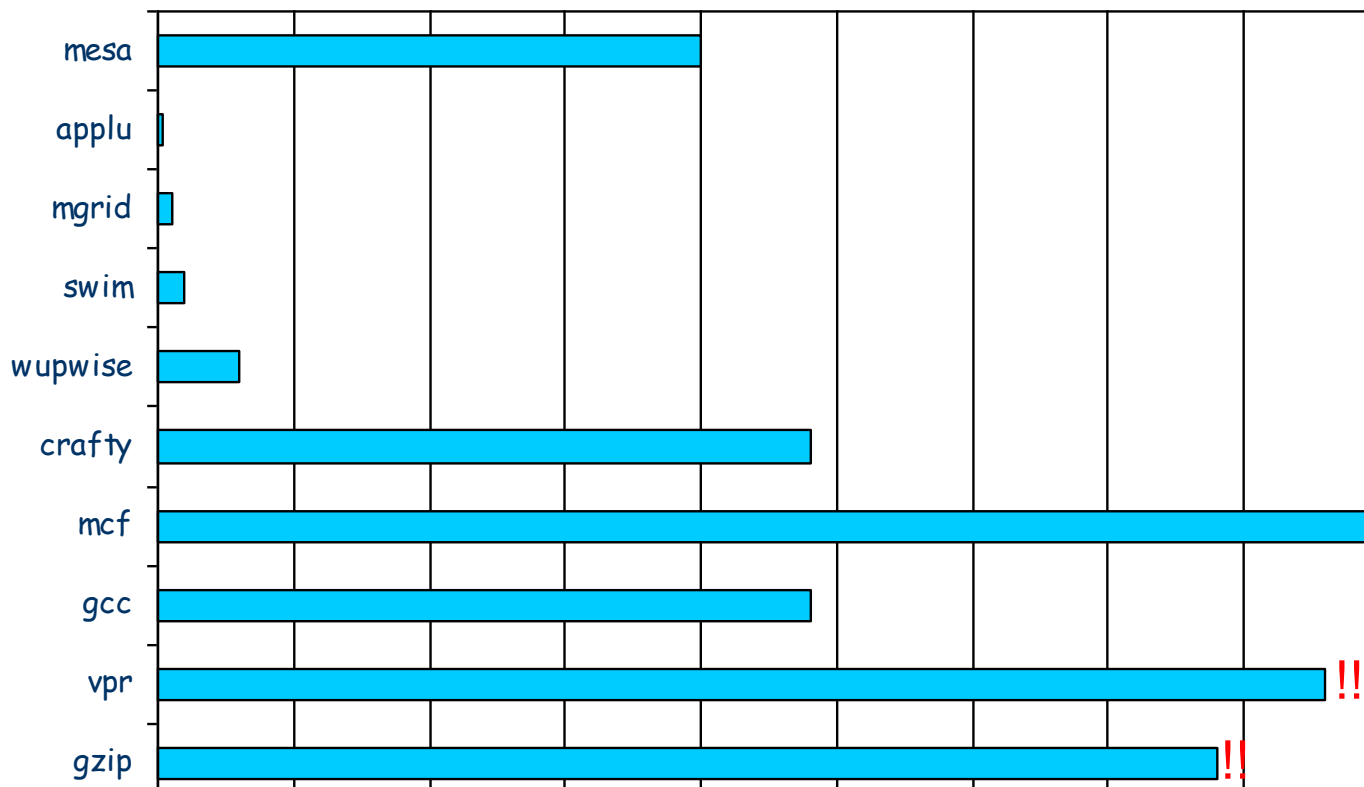
CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Netburst(Pentium4) Rendimiento

Especulación: porcentaje de μ -operaciones que no finalizan (commit)

Muy correlacionada con gráfica anterior



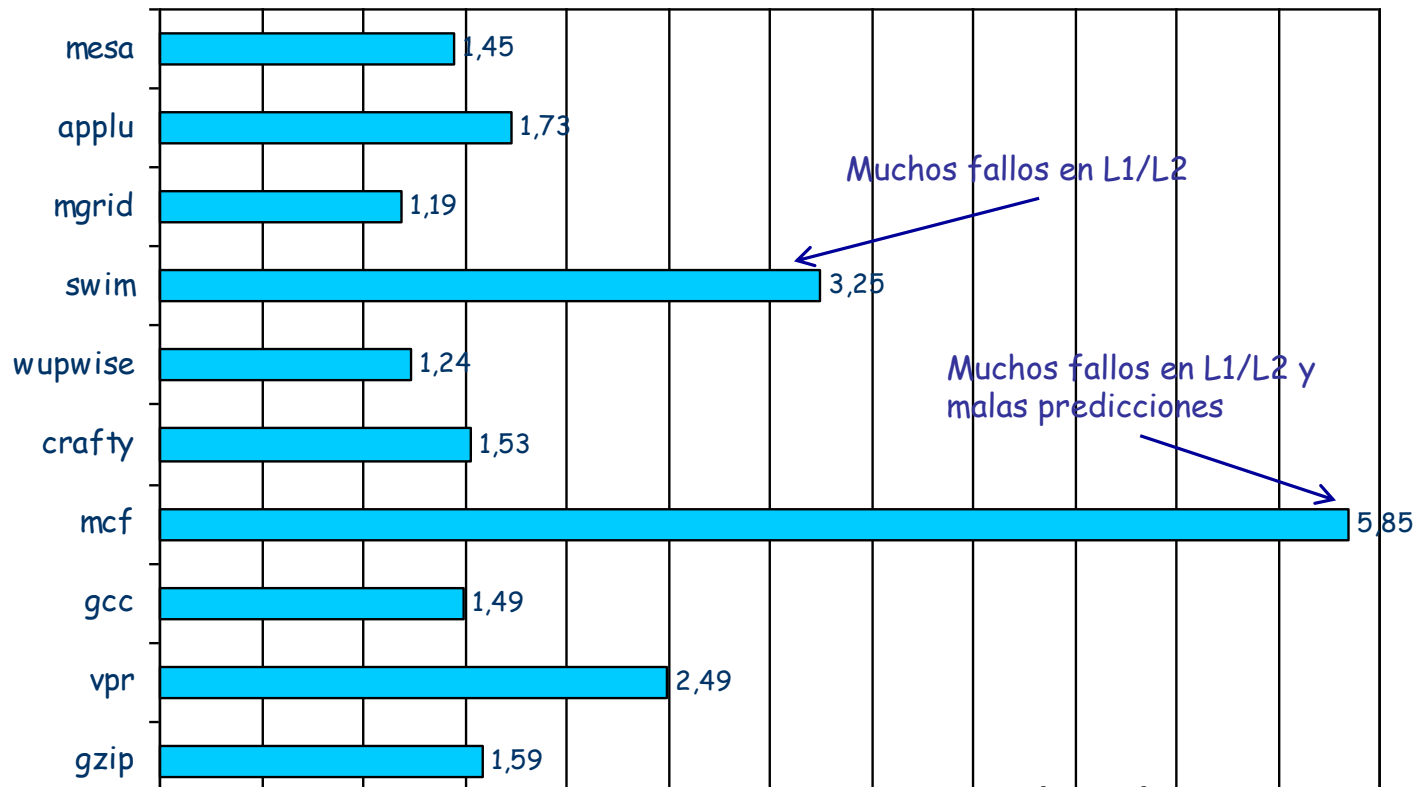
Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Netburst(Pentium4) Rendimiento

CPI real obtenido (ciclos promedio por instr IA-32)



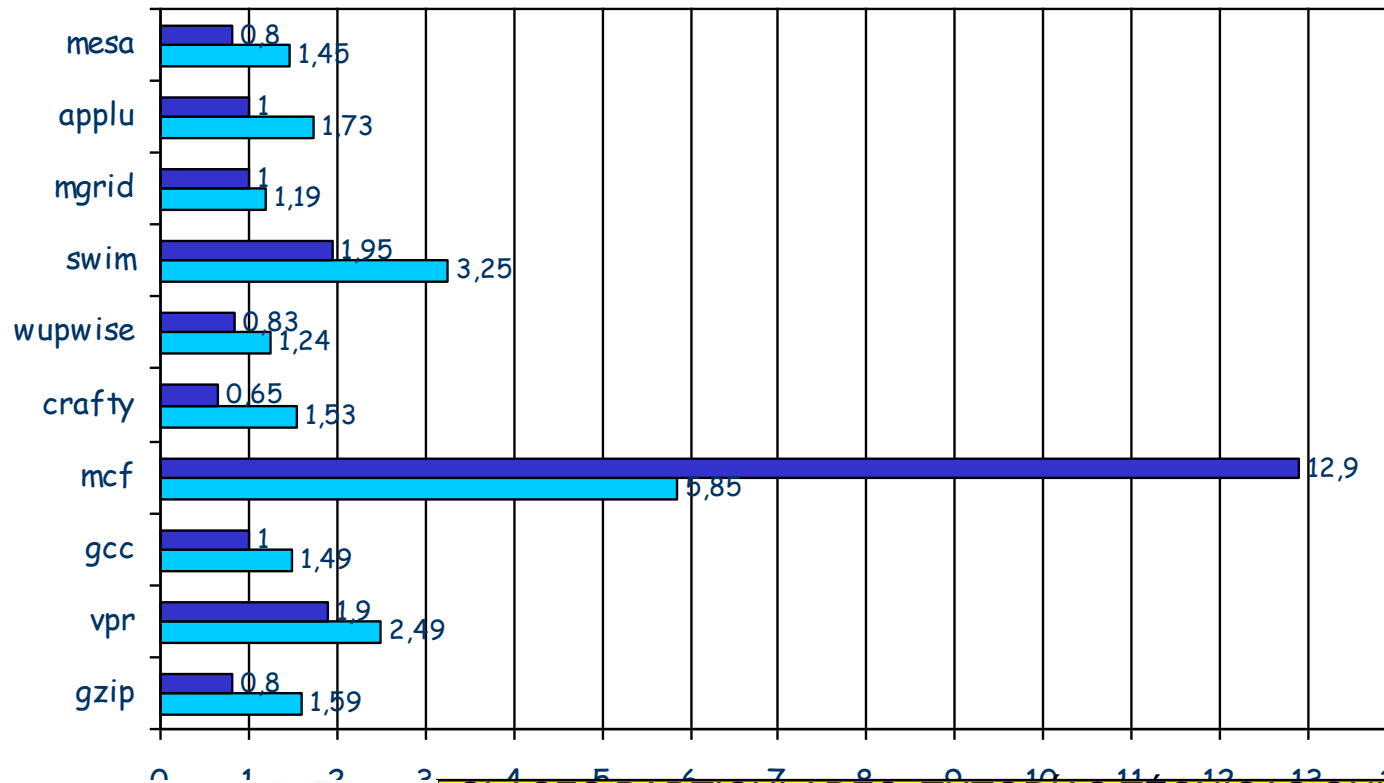
Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Netburst(Pentium4) versus AMD

Diferencia fundamental; Pentium pipe profundo para soportar alta frecuencia
CPI real obtenido



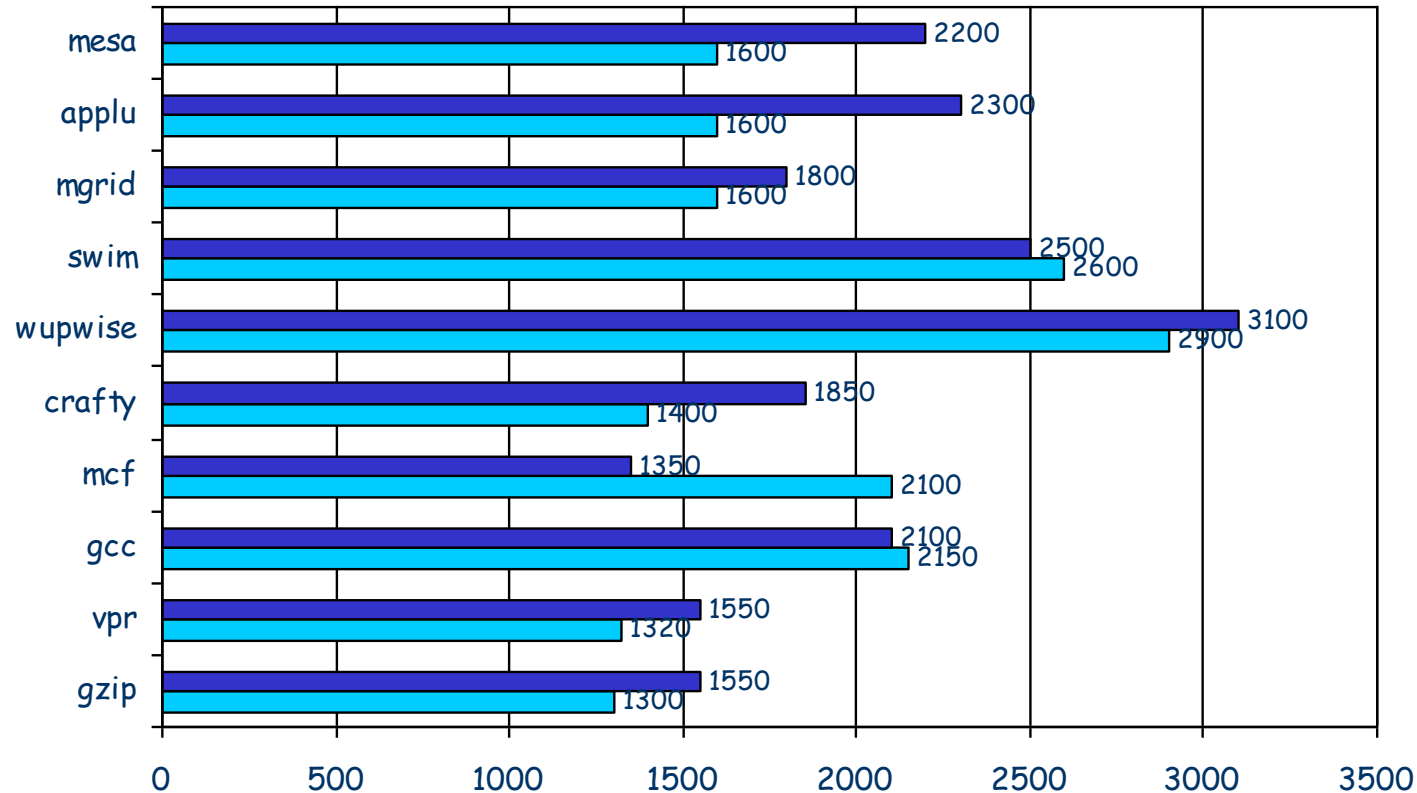
Cartagena99

CLASES PARTICULARES, TUTORIAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Netburst(Pentium4) versus AMD

Rendimiento en SPEC2000



Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

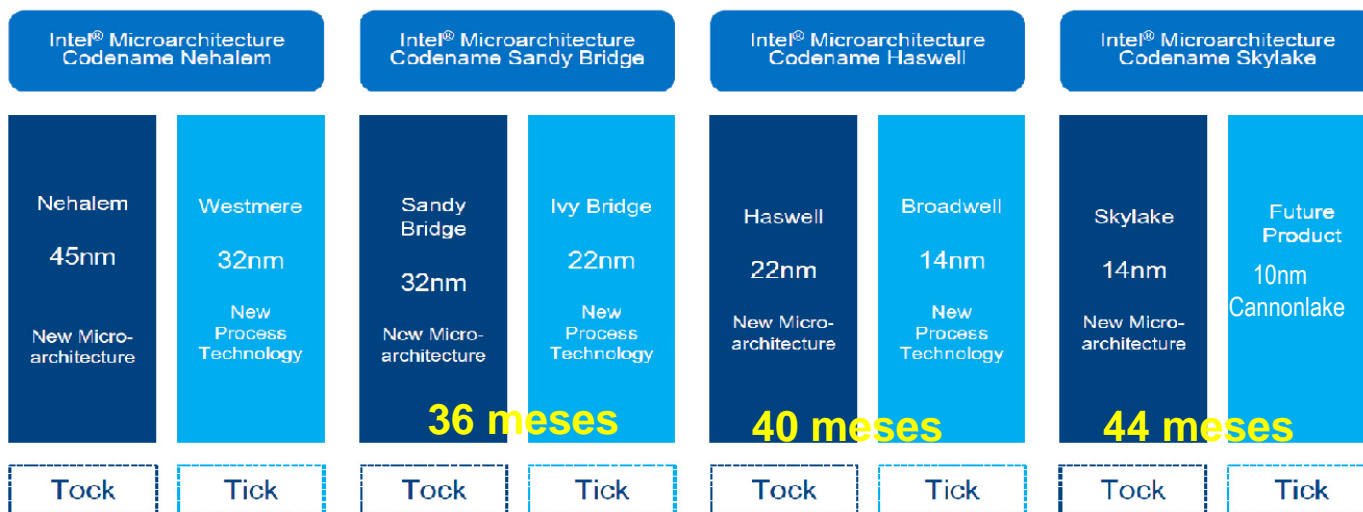
ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Intel Dual Core, Core2, Quad, Core i...

Tick-Tock

Tick-Tock Development Model:

Sustained Microprocessor Leadership



Innovation delivers new microarchitecture with Skylake

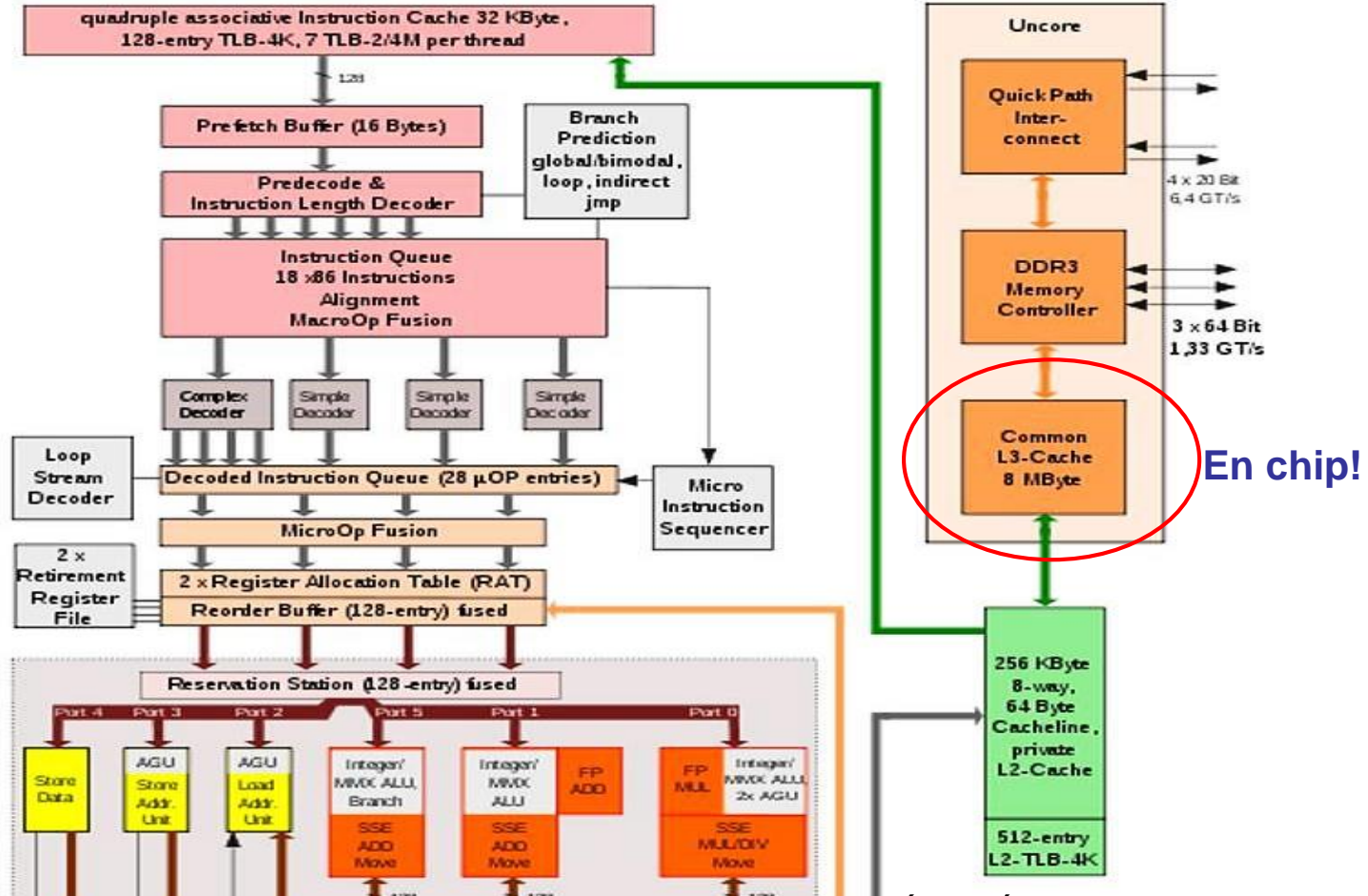
Intel
R&D PIPELINE

CLASES PARTICULARES, TUTORIAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Cartagena99

Nehalem Microarchitecture

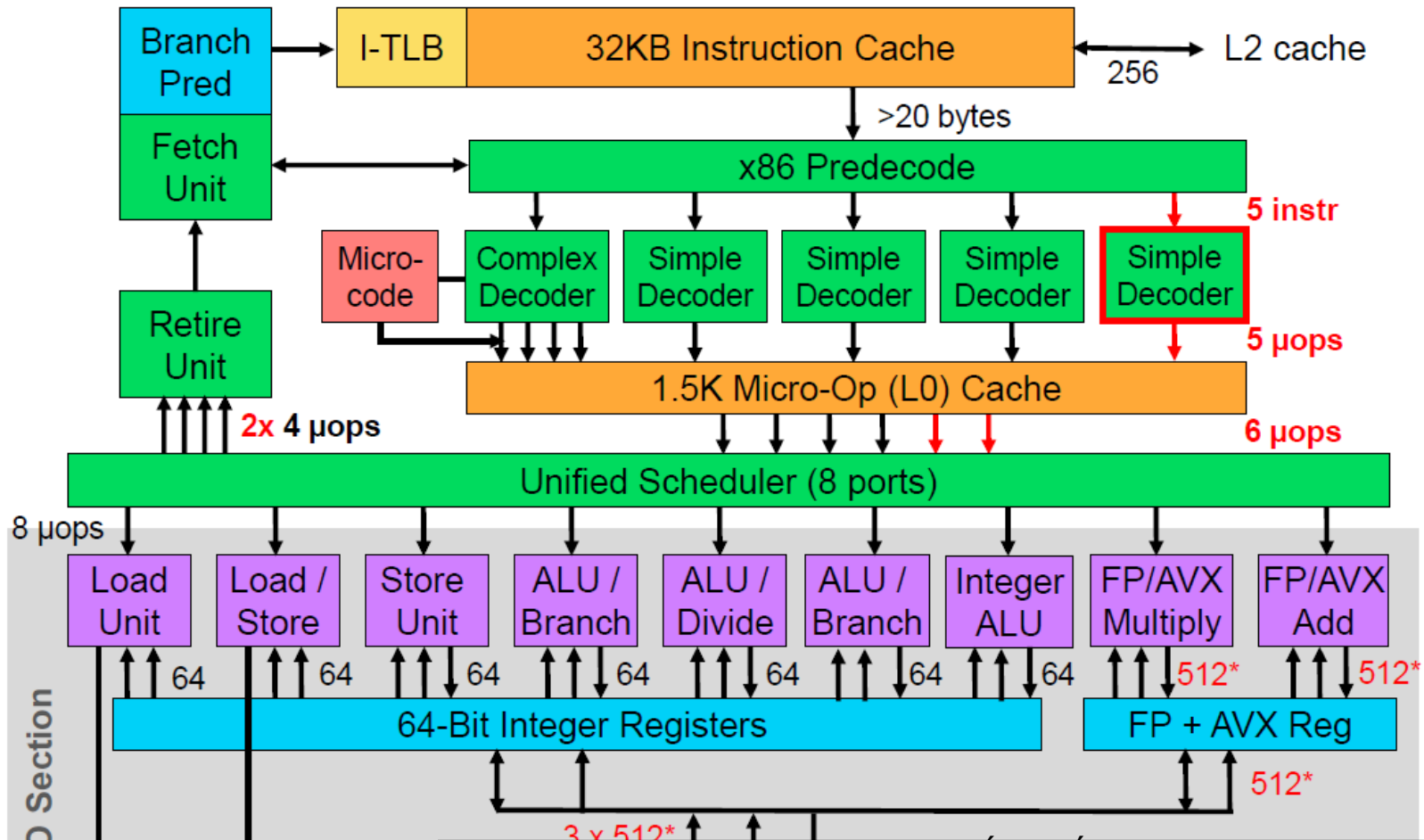


Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
 LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
 CALL OR WHATSAPP:689 45 44 70

Skylake Microarchitecture



CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
 LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
 CALL OR WHATSAPP:689 45 44 70

Cartagena99

Skylake Microarchitecture

	Nehalem	Sandy Bridge	Haswell	Skylake
x86 Decoders	4 instr	4 instr	4 instr	5 instr
Max Instr/Cycle	4 ops	6 ops	8 ops	8 ops
Reorder Buffer	128 ops	168 ops	192 ops	224 ops
Load Buffer	48 loads	64 loads	72 loads	72 loads
Store Buffer	32 stores	36 stores	42 stores	56 stores
Scheduler	36 entries	54 entries	60 entries	97 entries
Integer Rename	In ROB	160 regs	168 regs	180 regs
FP Rename	In ROB	144 regs	168 regs	168 regs
Allocation Queue	28/thread	28/thread	56 total	64/thread

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

- - -

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Límites del ILP

- ❑ ¿Como superar los limites de este estudio?
- ❑ Mejoras en los compiladores e ISAs
- ❑ Eliminar riesgos EDL y EDE en la memoria
- ❑ Eliminar riesgos LDE en registros y memoria. Predicción
- ❑ Buscar otros paralelismos (thread)

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Límites del ILP

❑ Buscar paralelismo de más de un thread

- o Hay mucho paralelismo en algunas aplicaciones (Bases de datos, códigos científicos)
- o Thread Level Parallelism
 - o Thread: proceso con sus propias instrucciones y datos
 - Cada thread puede ser parte de un programa paralelo de múltiples procesos, o un programa independiente.
 - Cada thread tiene todo el estado (instrucciones, datos, PC, register state, ...) necesario para permitir su ejecución
 - Arquitecturas(multiprocesadores, MultiThreading y multi/many cores)

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Thread Level Parallelism (TLP) versus ILP

- ❑ ILP explota paralelismo implícito dentro de un segmento de código lineal o un bucle
- ❑ TLP representa el uso de múltiples thread que son inherentemente paralelos.
- ❑ Objetivo: Usar múltiples streams de instrucciones para mejorar:
 - o Throughput de computadores que ejecutan muchos programas diferentes.
 - o Reducir el tiempo de ejecución de un programa multi-threaded

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Multithreading

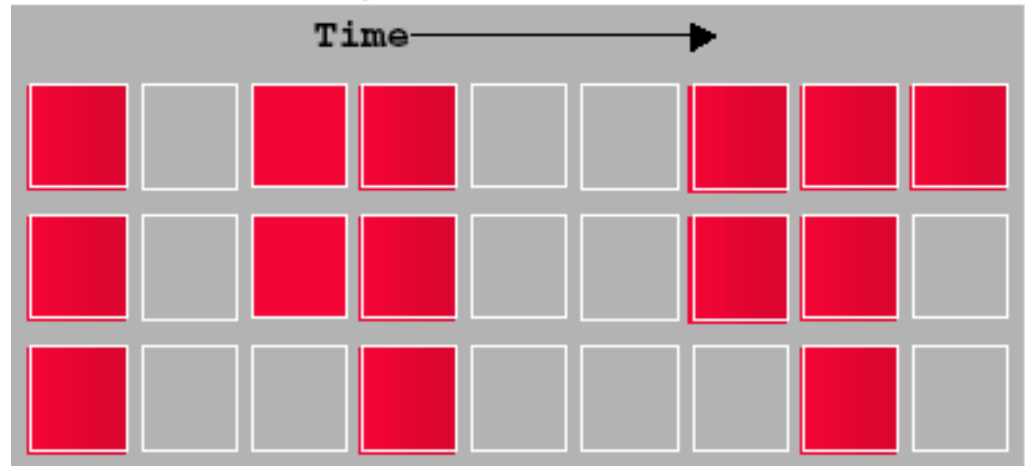
¿ Por que multithreading ?

❑ Procesador superescalar

(P. ej.: Hasta 3 Fus pueden comenzar una op en cada ciclo)

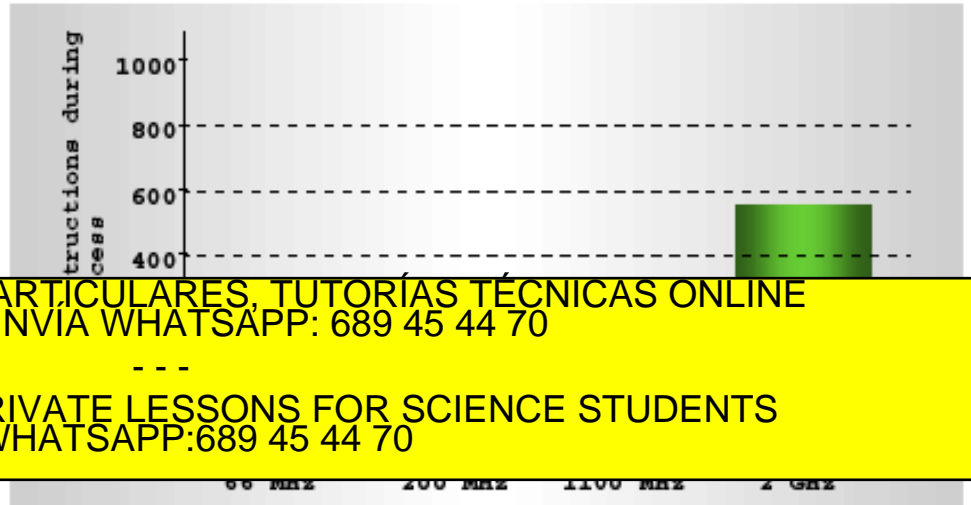
Causas de baja utilización de UFs:

- No hay instrucciones que ejecutar (fallo cache)
- Hay instr, pero faltan operandos
- Hay UFs libres, pero no del tipo necesario



❑ La latencia de memoria crece.

¿Como sobortarla?



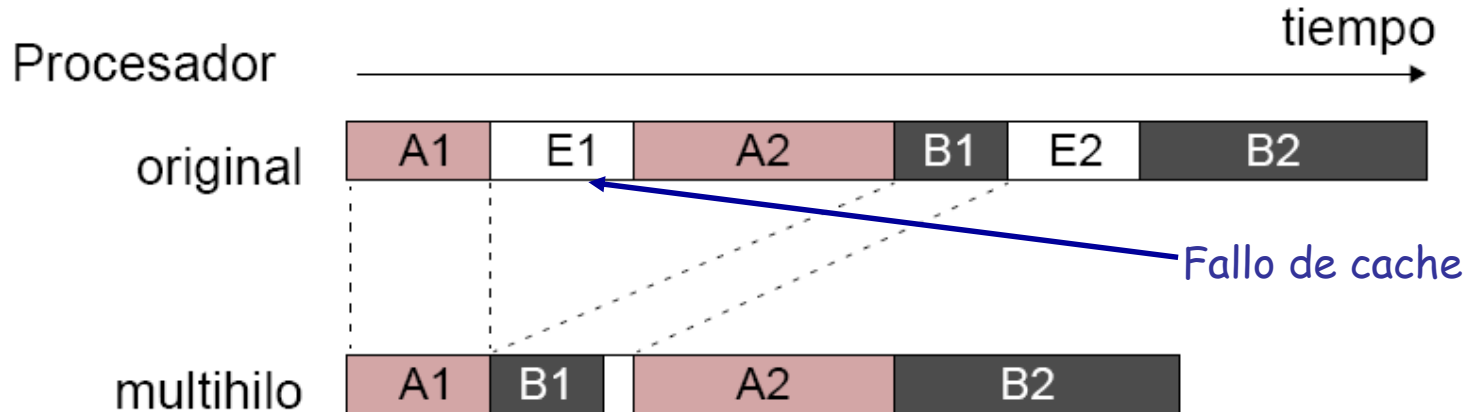
CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Cartagena99

Multithreading

□ Multithreading



- Incrementar el trabajo procesado por unidad de tiempo
- Si los hilos son del mismo trabajo se reduce el tiempo de ejecución

La técnica multithreading no es ideal y se producen pérdidas de rendimiento. Por ejemplo, un programa puede ver incrementado su tiempo de ejecución aunque el

Cartagena99

CLASES PARTICULARES, TUTORIAS TECNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Multithreading

□ Ejecución Multithreaded

- o Multithreading: múltiples threads comparten los recursos del procesador
 - o El procesador debe mantener el estado de cada thread e.g., una copia de bloque de registros, un PC separado, tablas de páginas separadas.
 - o La memoria compartida ya soporta múltiples procesos.
 - o HW para conmutación de thread muy rápido. Mucho mas rápido que entre procesos.
- o ¿Cuándo conmutar?
 - o Cada ciclo conmutar de thread (grano fino)
 - o Cuando un thread debe parar (por ejemplo fallo de cache)

Clave

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

□ Multithreading de Grano Fino

- o Conmuta entre threads en cada instrucción, entrelazando la ejecución de los diferentes thread.
- o Generalmente en modo "round-robin", los threads bloqueados se saltan
- o La CPU debe ser capaz de conmutar de thread cada ciclo.
- o Ventaja; puede ocultar stalls de alta y baja latencia, cuando un thread esta bloqueado los otros usan los recursos.
- o Desventaja; retarda la ejecución de cada thread individual, ya que un thread sin stall es retrasado por

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Multithreading

□ Multithreading Grano Grueso

- o Conmuta entre threads solo en caso de largos stalls, como fallos de cache L2
- o Ventajas
 - o No necesita conmutación entre thread muy rápida.
 - o No retarda cada thread, la conmutación solo se produce cuando un thread no puede avanzar.
- o Desventajas; no elimina perdidas por stalls cortos. La conmutación es costosa en ciclos.
 - o Como CPU lanza instrucciones de un nuevo thread, el pipeline debe ser vaciado.
 - o El nuevo thread debe llenar el pipe antes de que las

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Opinion 1.12

Multithreading

❑ Simultaneous Multithreading

Motivación: Recursos no usados en un procesador superescalares

Un thread, 8 unidades

Ciclo M M FX FX FP FP BR CC

1	█							█
2	█	█					█	
3			█	█				
4								
5								
6								
7	█			█		█		

Dos threads, 8 unidades

Ciclo M M FX FX FP FP BR CC

1	█	█	█					█
2	█	█	█			█	█	
3	█			█	█			
4	█	█				█		
5		█						█
6								
7	█		█	█	█	█		

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

M = Load/Store, FX = Fixed Point, FP = Floating Point, BR = Branch, CC = Condition Codes

Multithreading

□ Simultaneous Multithreading (SMT)

- Simultaneous multithreading (SMT): dentro de un procesador superescalar fuera de orden ya hay mecanismos Hw para soportar la ejecución de más de un thread
 - o Gran numero de registros físicos donde poder mapear los registros arquitectónicos de los diferentes threads
 - o El renombrado de registros proporciona un identificador único para los operandos de una instrucción, por tanto instrucciones de diferentes thread se pueden mezclar sin confundir sus operados
 - o La ejecución fuera de orden permite una utilización eficaz de los recursos.
- Solo necesitamos sumar una tabla de renombrado por thread y PC separados
 - o Commit independiente se soporta con un ROB por thread (Lógico o físico)
- Ojo conflictos en la jerarquía de memoria

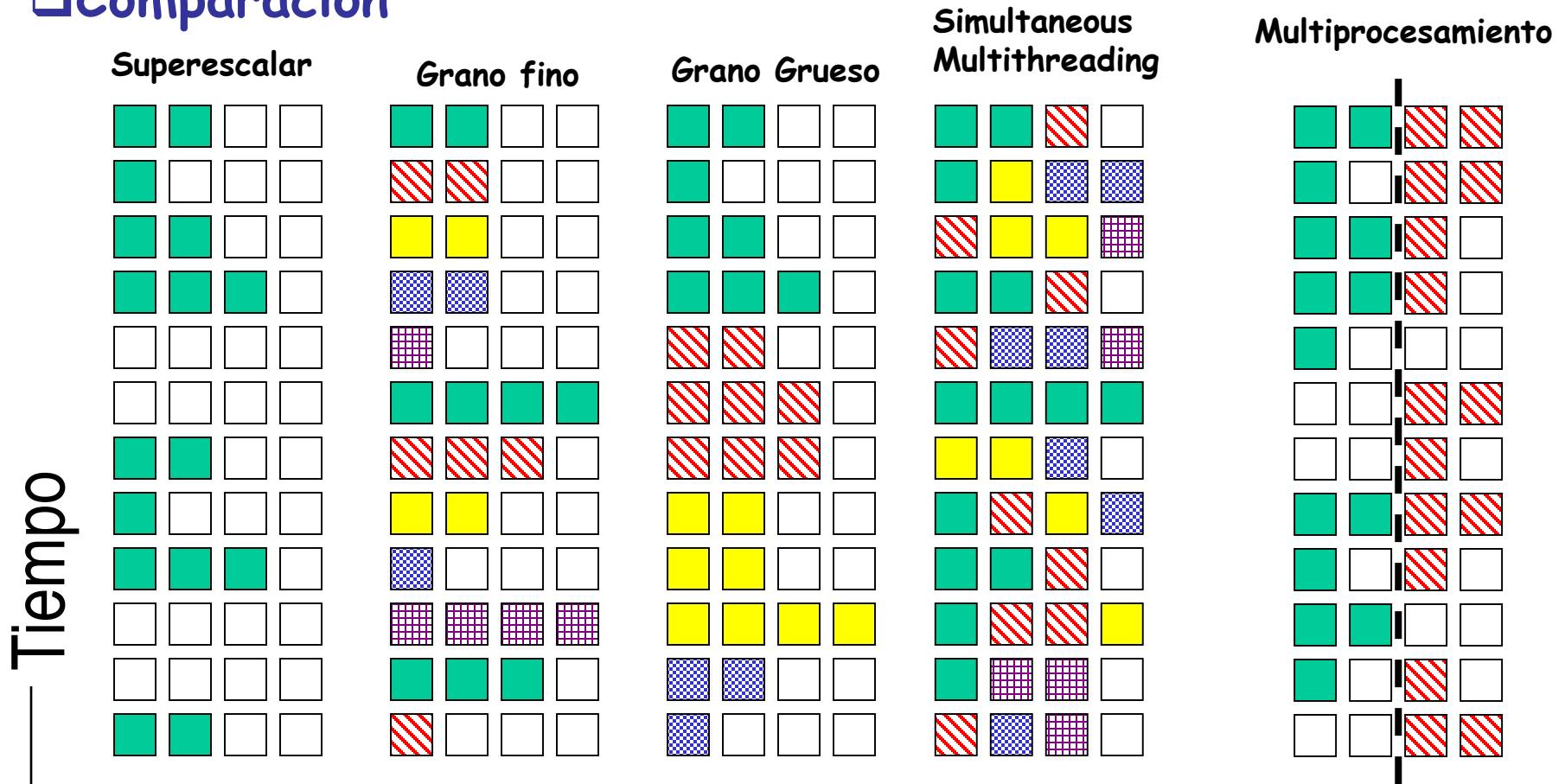
Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Multithreading

Comparación



Tiempo

Cartagena99

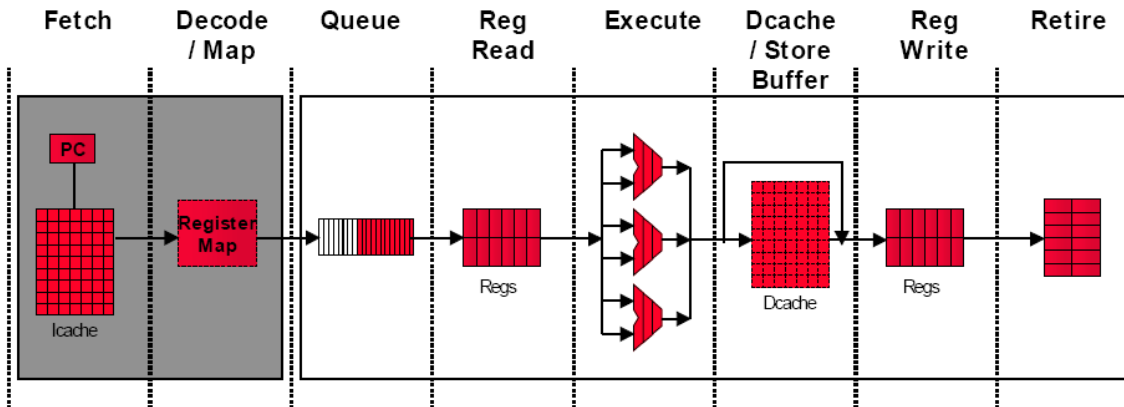
CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
 LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
 CALL OR WHATSAPP: 689 45 44 70

SMT Multithreading

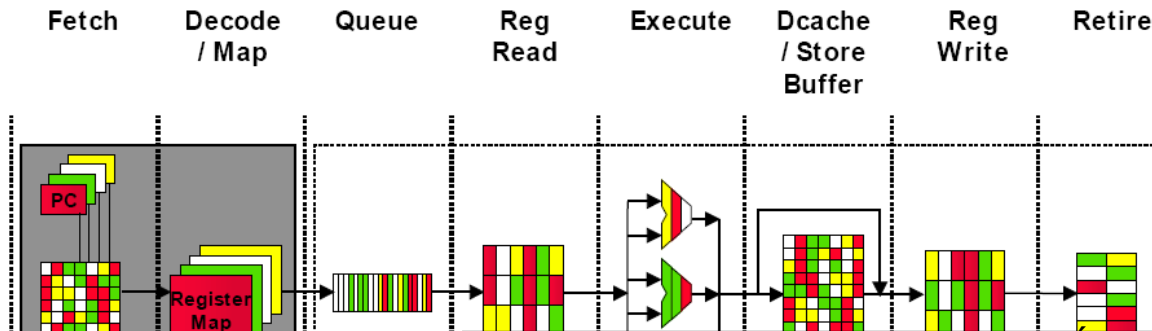
□ Simultaneous multithreading

- Un solo hilo: un flujo de instrucciones



✓ todos los recursos utilizados por un hilo

- Multihilo: varios flujos de instrucciones



CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

✓ los otros recursos se pueden compartir

Pentium-4 Hyperthreading (2002)

- ❑ Primer procesador comercial SMT 2 threads
- ❑ Los procesadores lógicos comparten casi todos los recursos del procesador físico.
 - o Caches, unidades de ejecución, predictor de saltos
- ❑ Overhead de área por hyperthreading ~ 5%
- ❑ Cuando un procesador lógico (thread) se detiene el otro puede progresar usando los recursos
 - o Ningun thread puede usar todos los recursos (colas) cuando hay 2 threads activos.
- ❑ Procesadores ejecutando solo un thread se ejecuta a la misma velocidad que sin hyperthreading

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

- - -

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

SMT Multithreading

❑ Power 4 (IBM 2000)

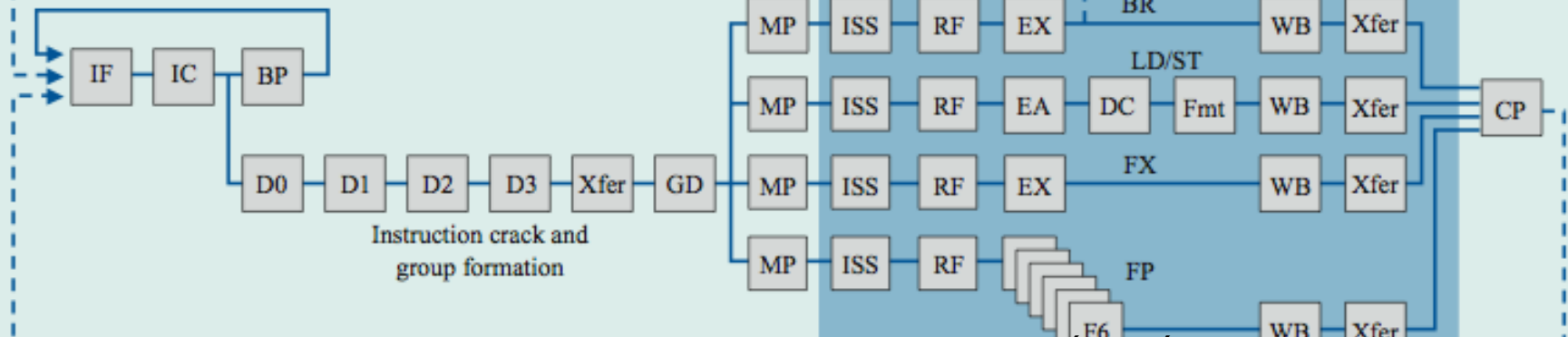
Predecesor Single-threaded del Power 5.
8 unidades de ejecución fuera de orden



Branch redirects

Instruction fetch

Out-of-order processing

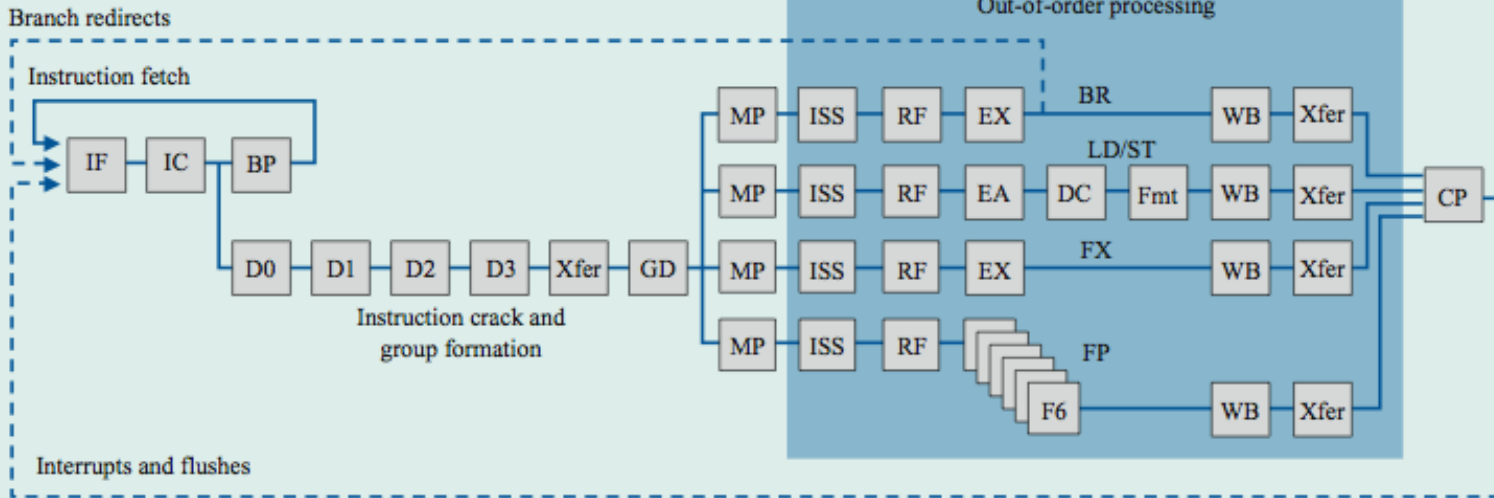


Cartagena99

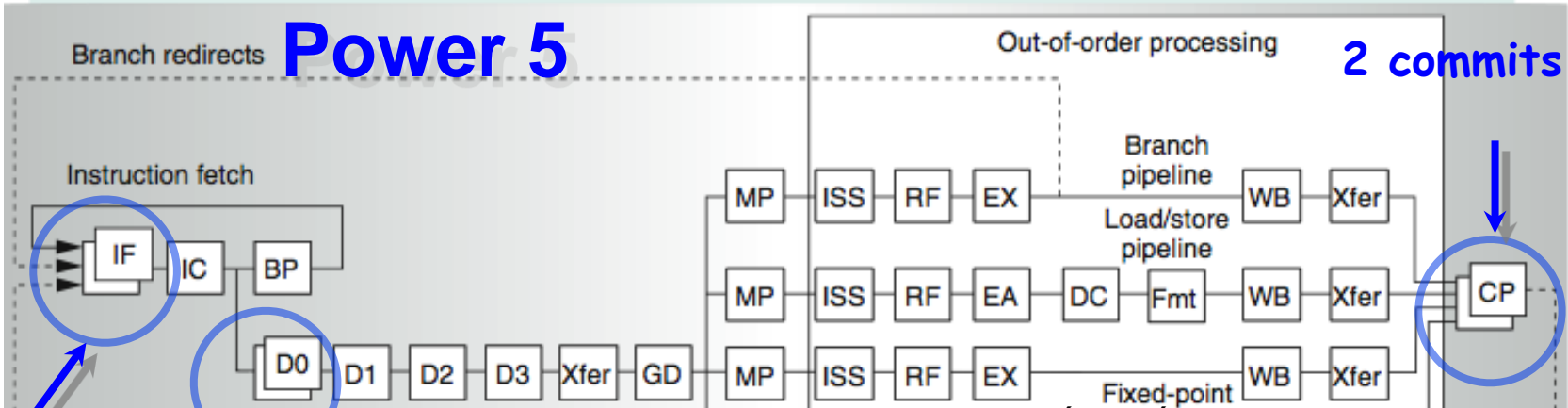
CLASES PARTICULARES, TUTORIAS TÉCNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Power 4



Power 5



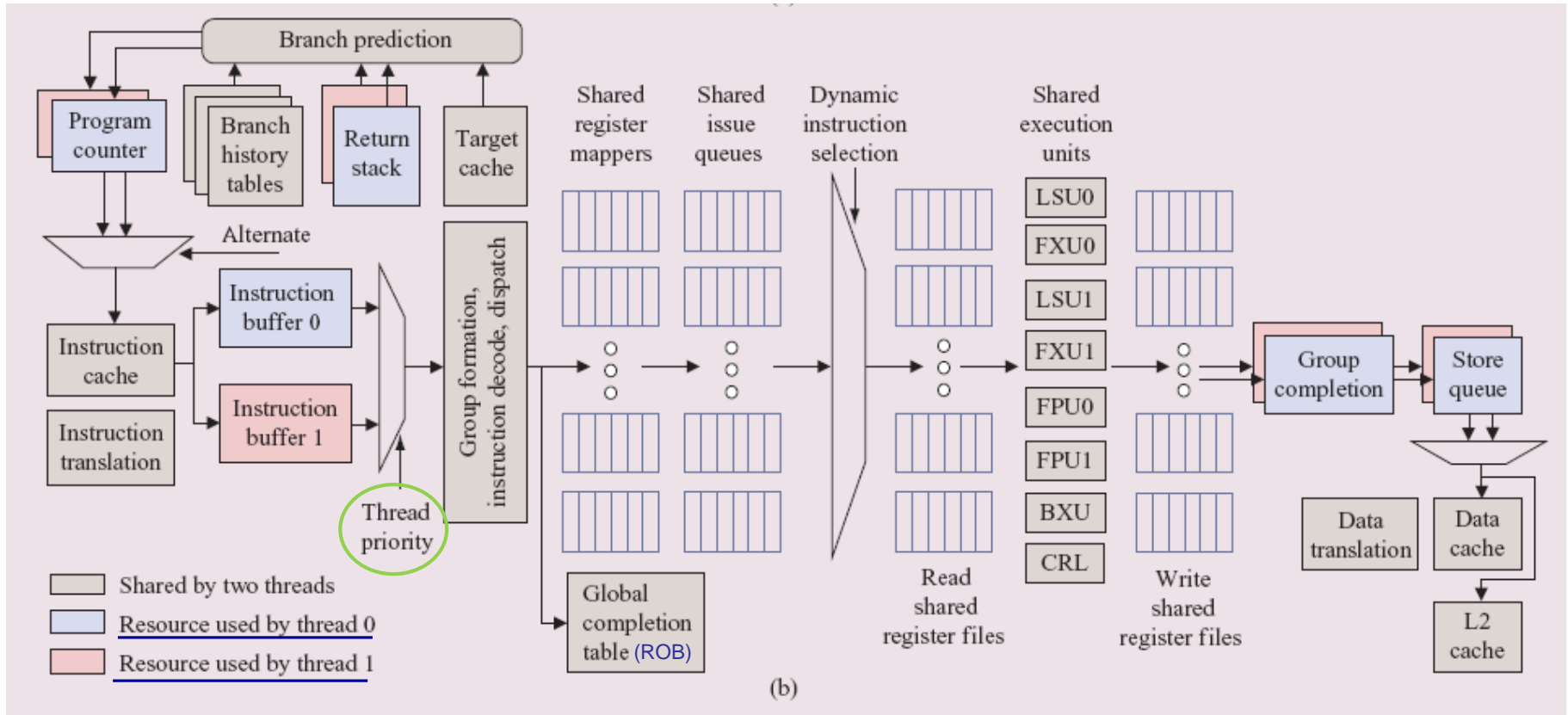
Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
 LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
 CALL OR WHATSAPP: 689 45 44 70

SMT Multithreading

□ Power 5 (IBM 2005)



Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP: 689 45 44 70

SMT Multithreading

□ Power 5

Balanced dynamic load

1- Monitorizar

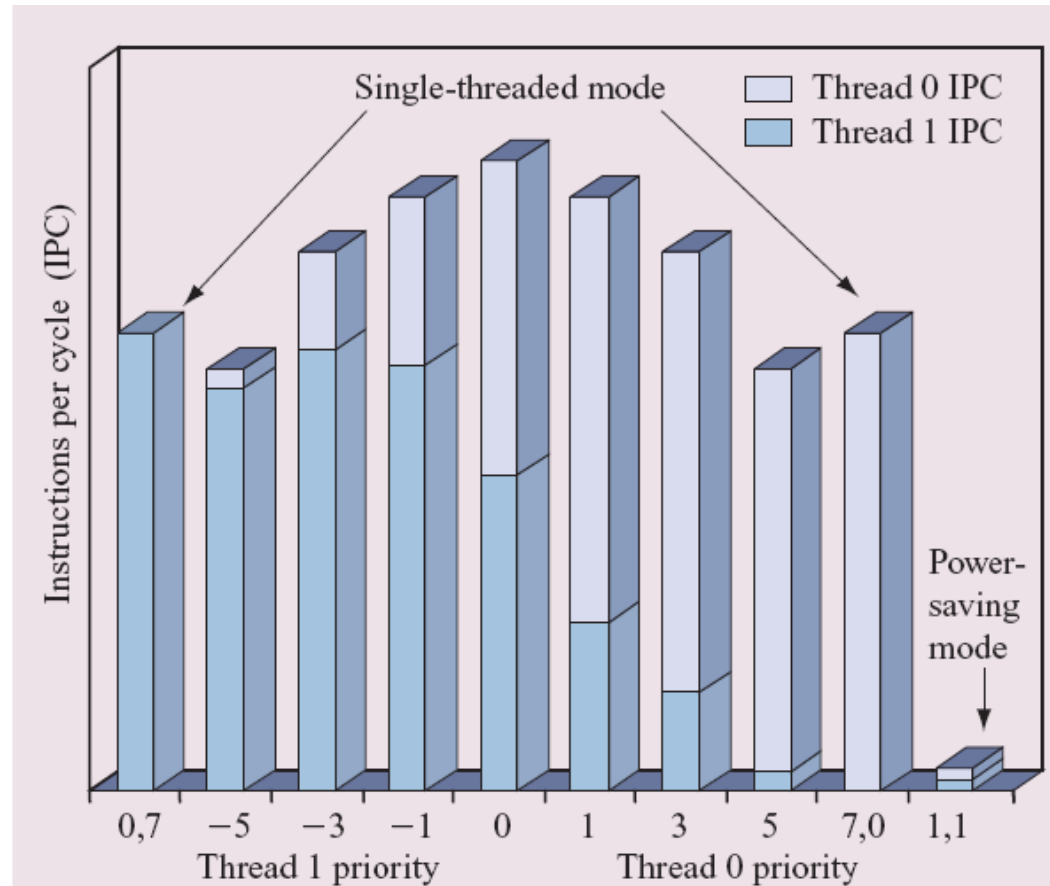
cola de fallos (load en L2)
entradas ocupadas en ROB (GCT)

2 - Quitarle recursos;

Reducir prioridad del hilo,
Inhibir decodificación (L2 miss)
Eliminar instrucciones desde emisión y
parar decodificación

3- Ajustar prioridad del hilo (hardware/software)

Baja en espera activa



Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP: 689 45 44 70

SMT Multithreading

- ❑ Cambios en Power 5 para soportar SMT
 - ❑ Incrementar asociatividad de la L1 de instrucciones y del TLB
 - ❑ Una cola de load/stores por thread
 - ❑ Incremento de tamaño de la L2 (1.92 vs. 1.44 MB) y L3
 - ❑ Un buffer de prebúsqueda separado por thread
 - ❑ Incrementar el número de registros físicos de 152 a 240
 - ❑ Incrementar el tamaño de la colas de emisión
 - ❑ El Power5 core es 24% mayor que el del Power4 para soportar SMT

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

SMT Multithreading

POWER 8

Technology

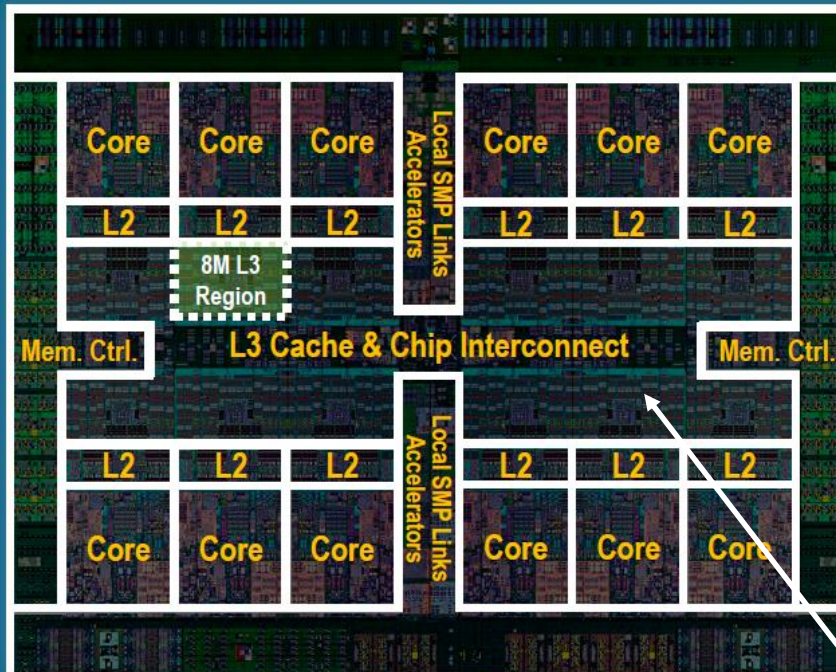
- 22nm SOI, eDRAM, 15 ML 650mm²

Cores

- 12 cores (SMT8)
- 8 dispatch, 10 issue, 16 exec pipe
- 2X internal data flows/queues
- Enhanced prefetching
- 64K data cache, 32K instruction cache

Accelerators

- Crypto & memory expansion
- Transactional Memory
- VMM assist
- Data Move / VM Mobility



Caches

- 512 KB SRAM L2 / core
- 96 MB eDRAM shared L3
- Up to 128 MB eDRAM L4 (off-chip)

Memory

- Up to 230 GB/s sustained bandwidth

Bus Interfaces

- Durable open memory attach interface
- Integrated PCIe Gen3
- SMP Interconnect
- CAPI (Coherent Accelerator Processor)

Energy Management

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

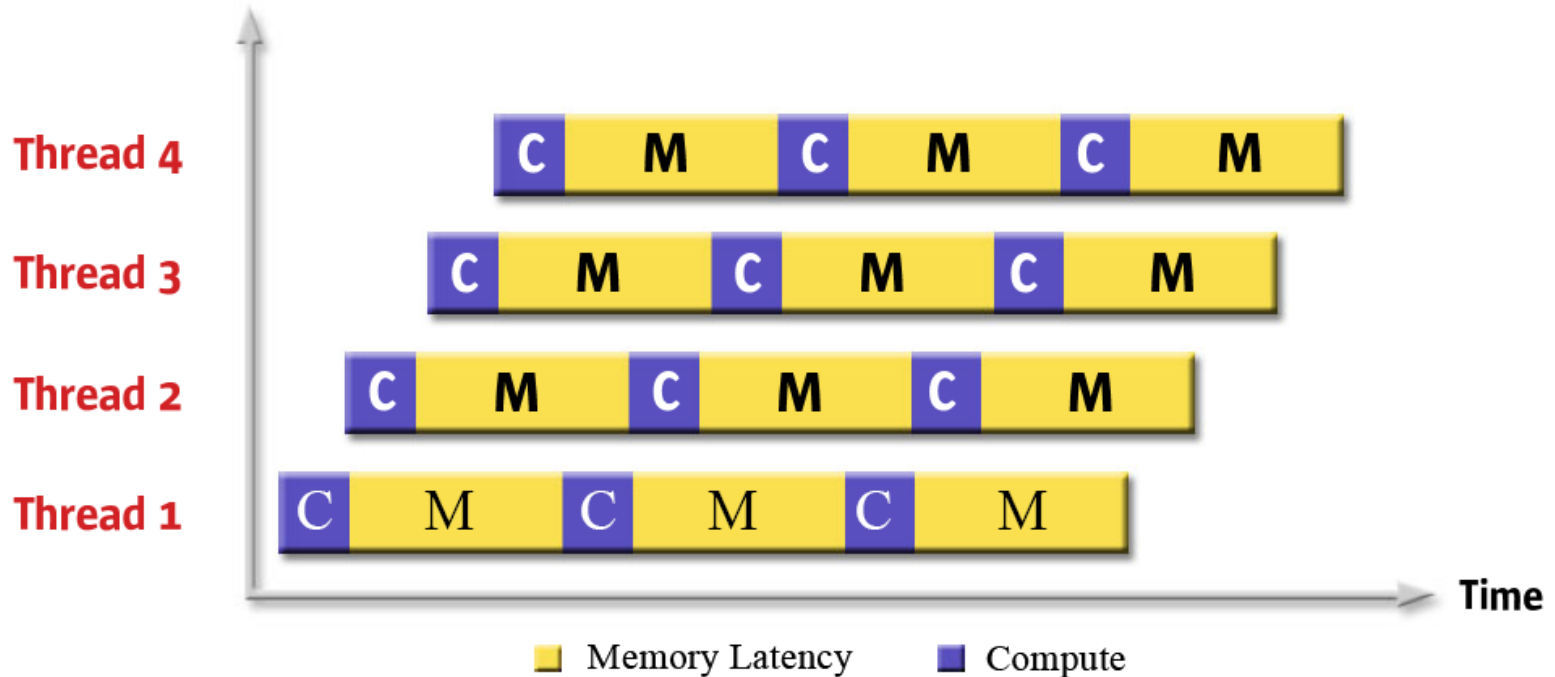
uso de embedded DRAM

Cartagena99

Multiprocesador en un Chip+ Multithreading grano fino

❑ Niagara (SUN 2005)

- ❑ Tolerar (soportar) la latencia de memoria mediante hilos concurrentes

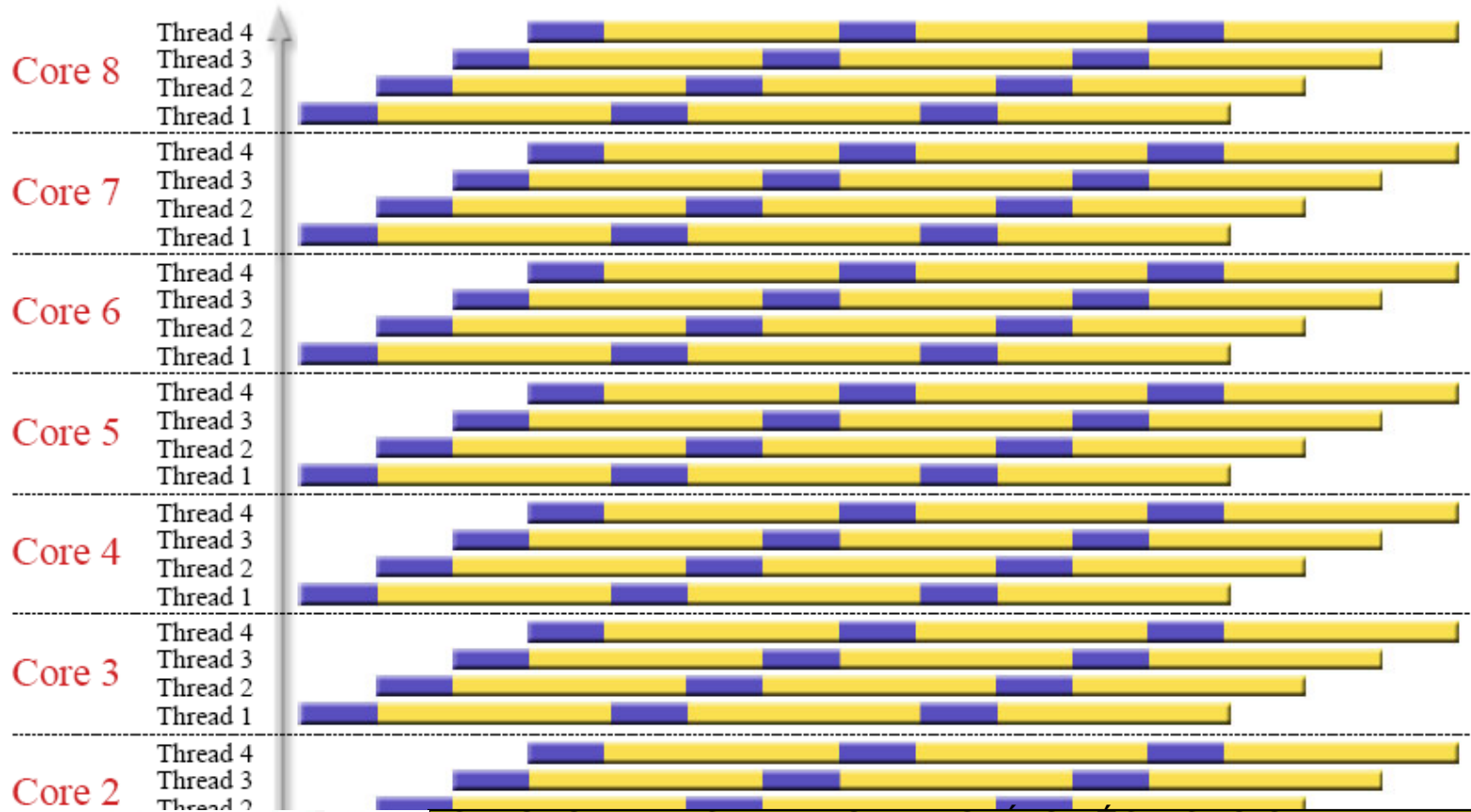


Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

❑ Niagara: Múltiples cores-múltiples thread



CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

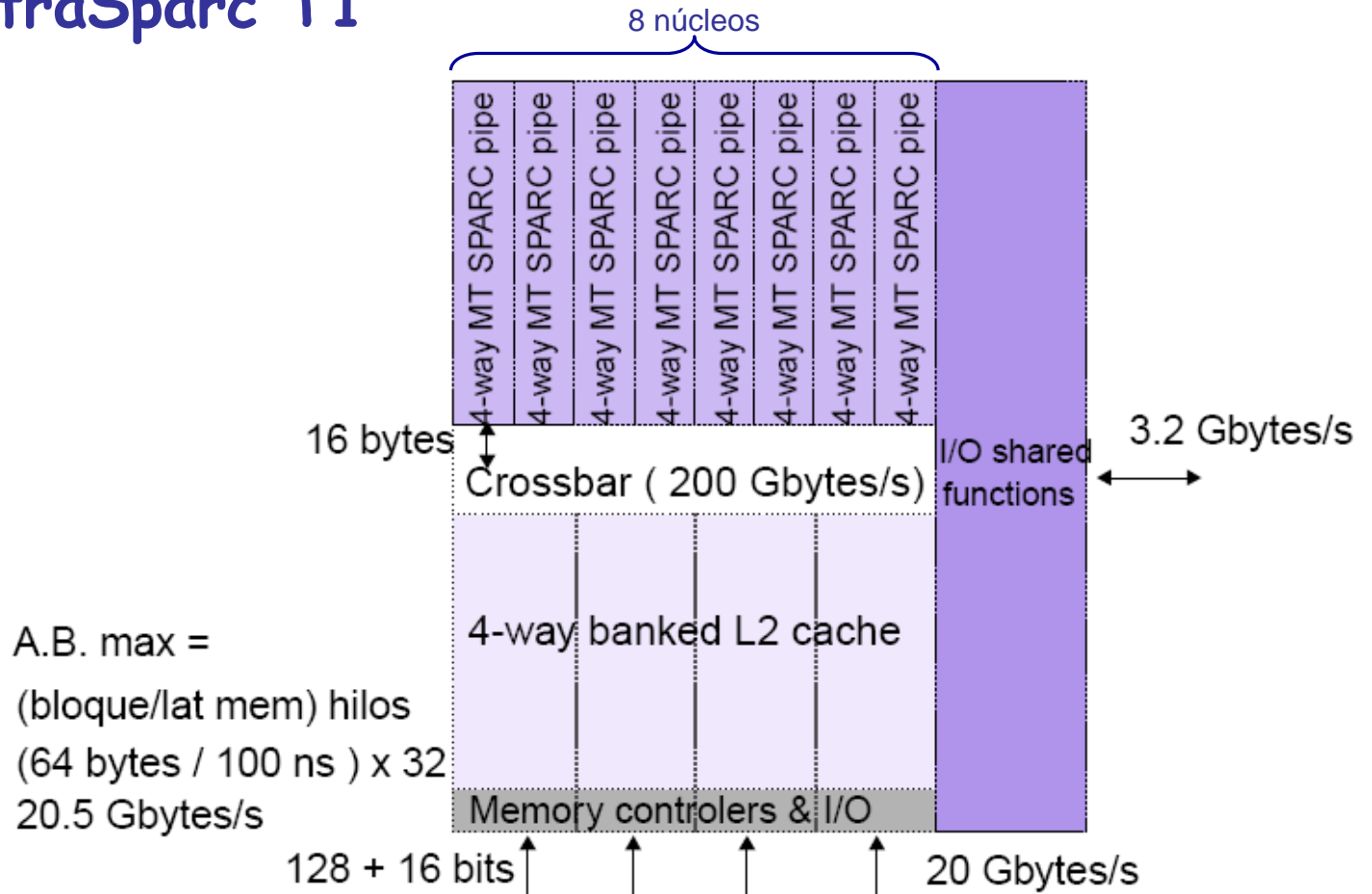
ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

■ Memory Latency ■ Compute

Cartagena99

Multiprocesador en un Chip+ Multithreading grano fino

❑ Niagara UltraSparc T1



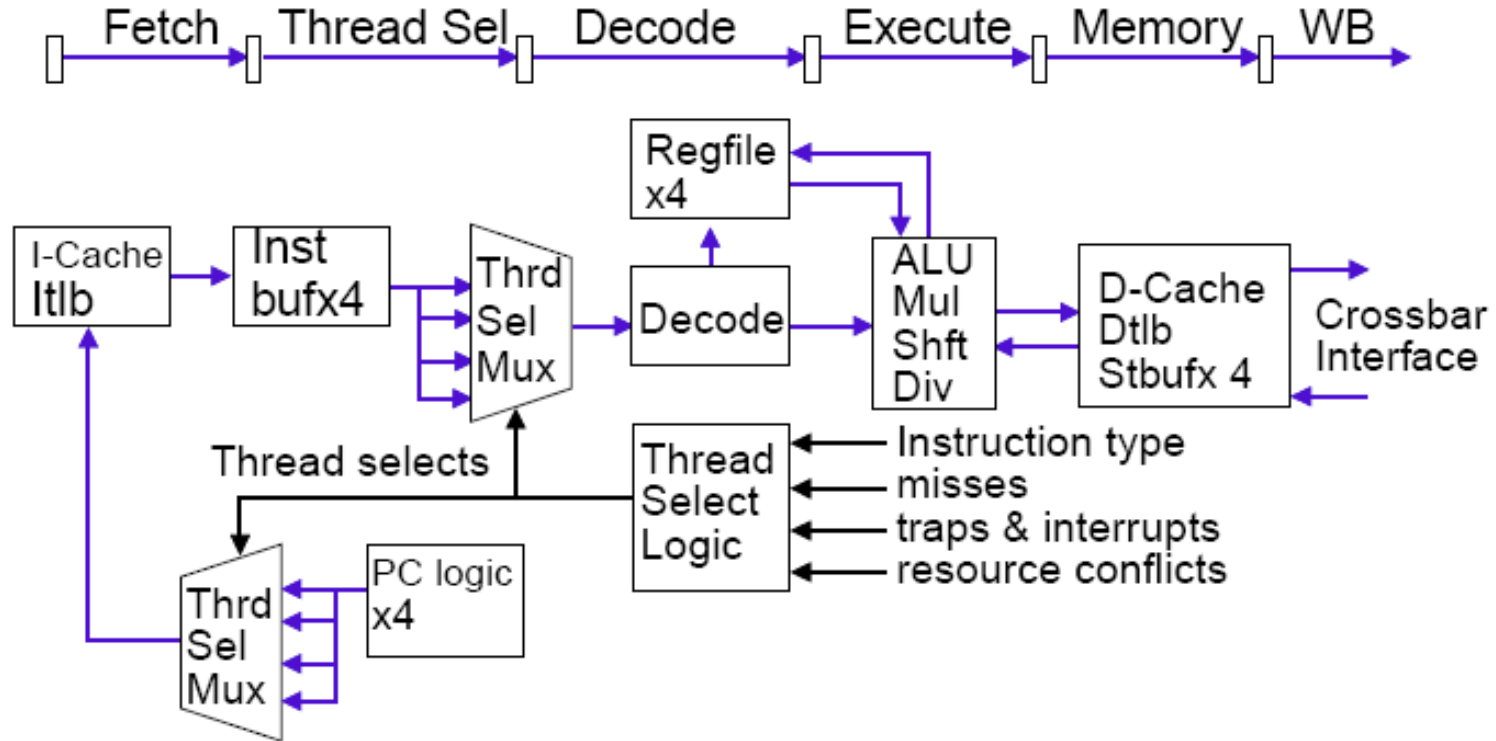
Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
 LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
 CALL OR WHATSAPP:689 45 44 70

Crossbar no bloqueante, No predictor de saltos, no FP (1 por chip)

❑ Niagara UltraSparcT1



- 6 etapas
- 4 thread independientes

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

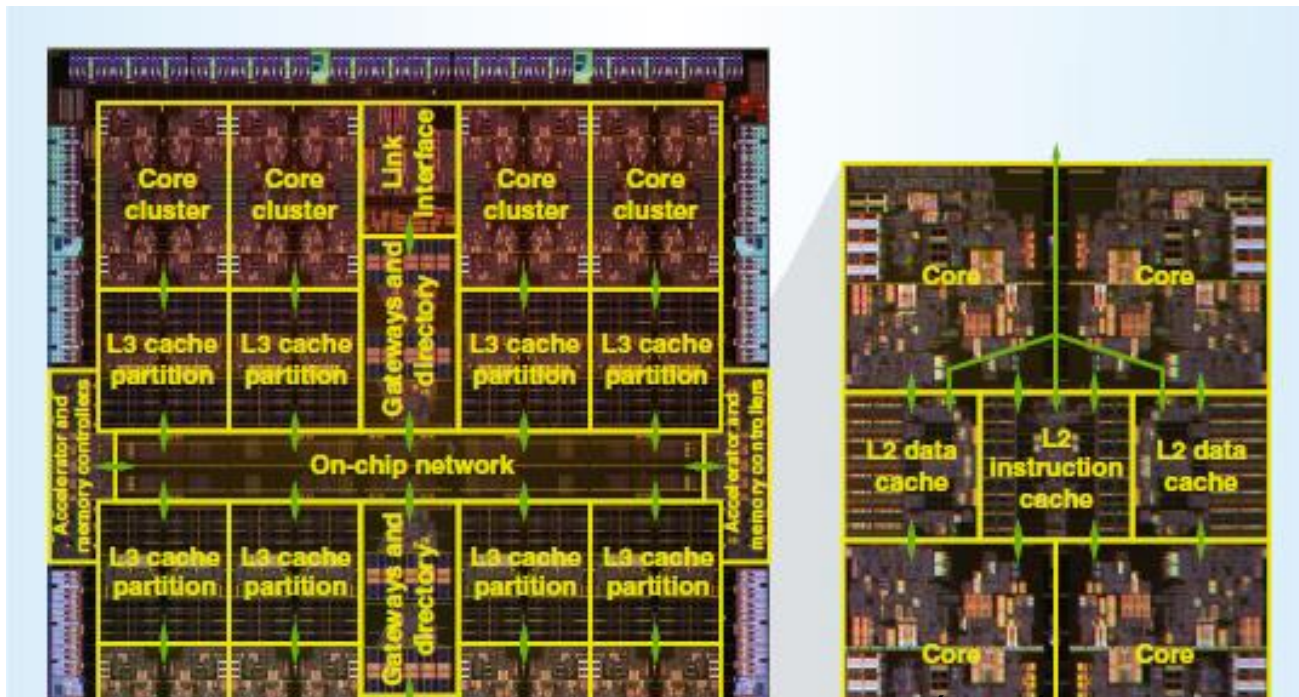
ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

▼ cada ciclo elige un hilo

Evolución Sparc

❑ Oracle Sparc M7

- o 32 cores / 8 threads, 256 threads/chip, 28nm, 10000 Mtrans, 699mm², 3.8 Ghz, 70MB L3
- o Hasta 8 socket/nodo: 2048 threads



Cartagena99

CLASES PARTICULARES, TUTORIAS TECNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Procesadores Oracle/Sun Niagara

- ❑ Orientación datacenters ejecutando servicios web y bases de datos (Oracle), con muchos threads ligeros.
- ❑ Muchos cores simples con reducida energía/ operación. Bajo rendimiento por thread

- ❑ Niagara-1 [2004], 8 cores, 4 threads/core
- ❑ Niagara-2 [2007], 8 cores, 8 threads/core
- ❑ Niagara-3 [2009], 16 cores, 8 threads/core
- ❑ T4 [2011], 8 cores, 8 threads/core
- ❑ T5 [2012], 16 cores, 8 threads/core
- ❑ M5 [2012], 6 cores, 8 threads/core
- ❑ M6[2013] 12 cores, 8 threads/core
- ❑ M7 [2014], 32 cores, 8 threads/core

	Sparc M7	Sparc M6	Sparc M5
CPU's	32 CPU's	12 CPU's	6 CPU's
CPU Type	SPARC V9 S4	SPARC V9 S3	SPARC V9 S3
CPU Freq (max)	3.8GHz*	3.6GHz	3.6GHz
CPU Threads	8 threads per CPU, 256 threads total	8 threads per CPU, 96 threads total	8 threads per CPU, 48 threads total
CPU Cluster	4 CPU's per cluster	CPU's not clustered	CPU's not clustered
Superscalar Issue	Dual issue, out of order		
Pipeline Depth	16 stages		
L1 Cache (I+D)	16KB + 16KB		
L2 Cache	256KB I-cache per cluster, 256KB D-cache per pair, 6MB total	128KB per CPU, 1.5MB total	128KB per CPU, 768KB total
L3 Cache	8MB per cluster, 64MB total	48MB L3	48MB L3
DRAM Interface	16x 64-bit DDR4-2133/2400/2667	8x 64-bit DDR3-1066	8x 64-bit DDR3-1066
DRAM Bandwidth (peak)	341.3GB/s	68.2GB/s	68.2GB/s
DRAM Bandwidth (measured)	170GB/s (DDR4-2133)	57GB/s (DDR3-1066)	57GB/s (DDR3-1066)
DRAM Capacity	2TB per socket	1TB per socket	1TB per socket
Glueless SMP	Up to 8 sockets		
Coherent SMP	Up to 32 sockets	Up to 96 sockets	Up to 32 sockets

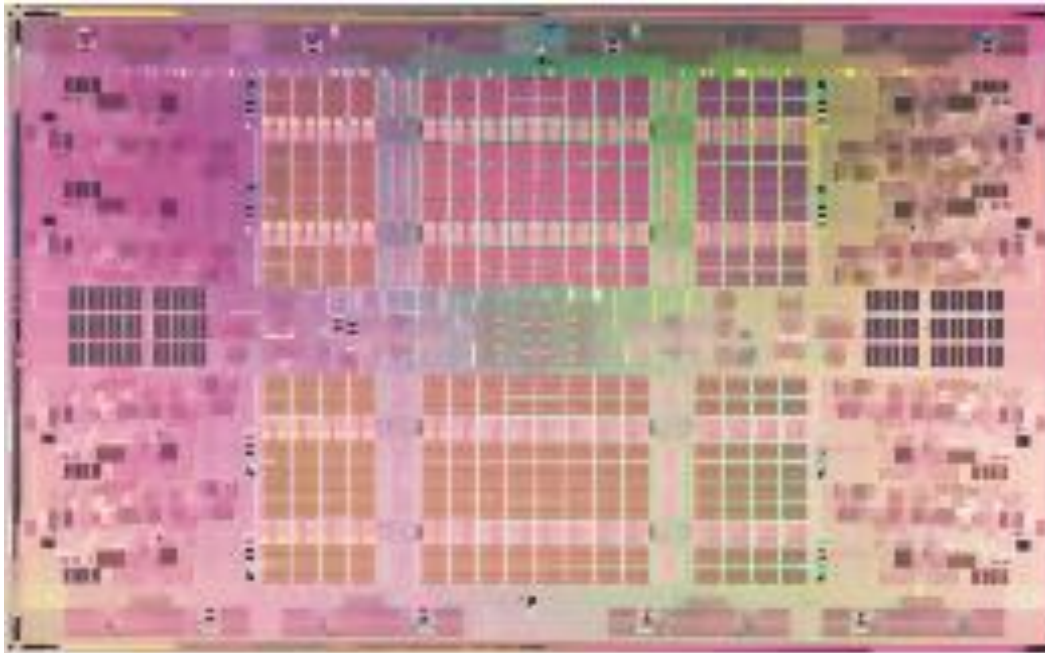
Cartagena99

CLASES PARTICULARES, TUTORIAS TECNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

VLIW EPIC-IA 64+ Multithreading grano grueso

□ Itanium2 9000 Poulson 2011



8 cores y 2 threads/core, 544mm², 3000MTs

Cache L2 separada , 512 KByte L2 Icache, 256KB L2 Dcache

Cache L3 32MB compartida

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

¿ Quien es mejor?

Procesador	Microarquitectura	Fetch / Issue / Execute	FU	Clock Rate (GHz)	Transistores Die size	Power
Intel Pentium 4 Extreme	Especulativo con planificación dinámica; Pipe profundo; SMT	3/3/4	7 int. 1 FP	3.8	125 M 122 mm ²	115 W
AMD Athlon 64 FX-57	Especulativo con planificación dinámica.	3/3/4	6 int. 3 FP	2.8	114 M 115 mm ²	104 W
IBM Power5 (1 CPU only)	Especulativo con planificación dinámica; SMT 2 CPU cores/chip	8/4/8	6 int. 2 FP	1.9	200 M 300 mm ² (est.)	80W (est.)
Intel	Planificación estática	6/5/11	9 int. 2	1.6	592 M	130 W



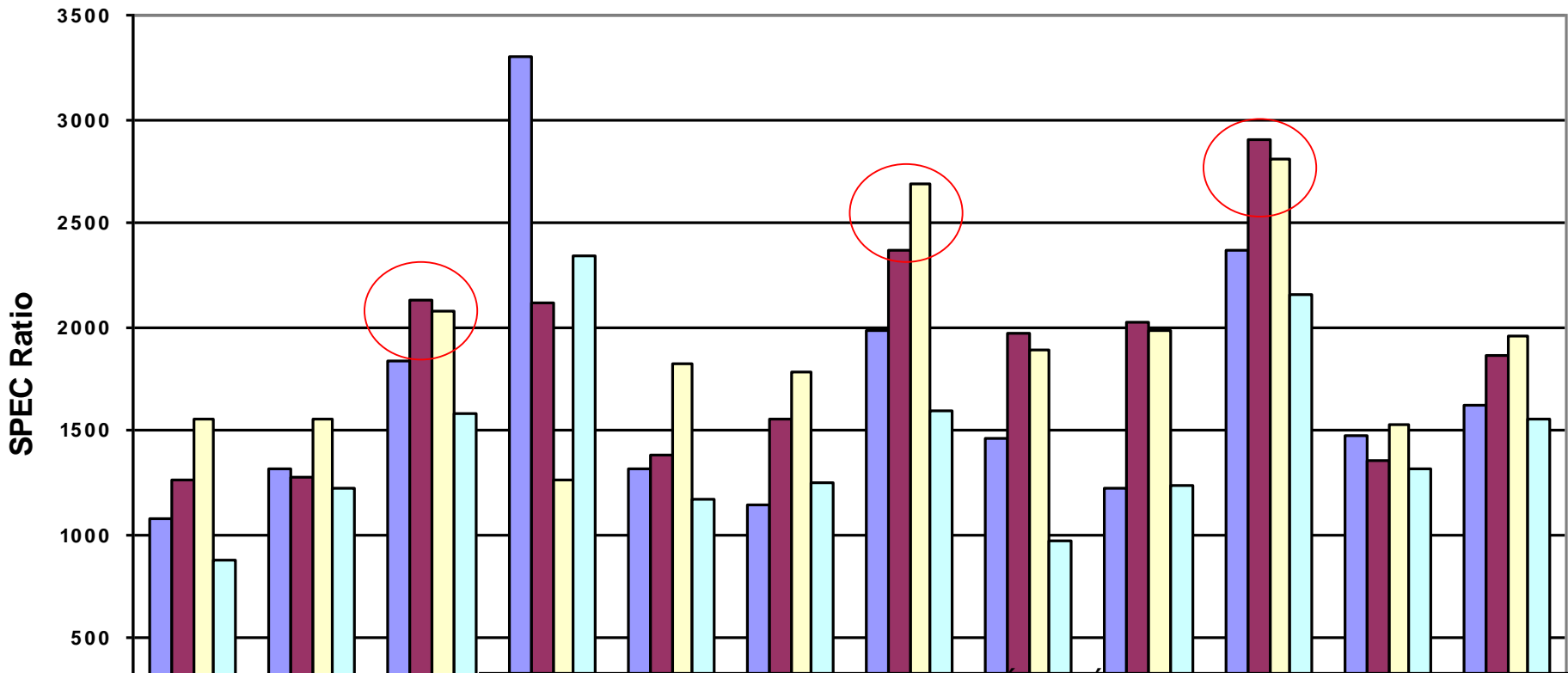
CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Rendimiento

❑ SPECint2000

■ Itanium 2 ■ Pentium 4 □ AMD Athlon 64 □ Power 5



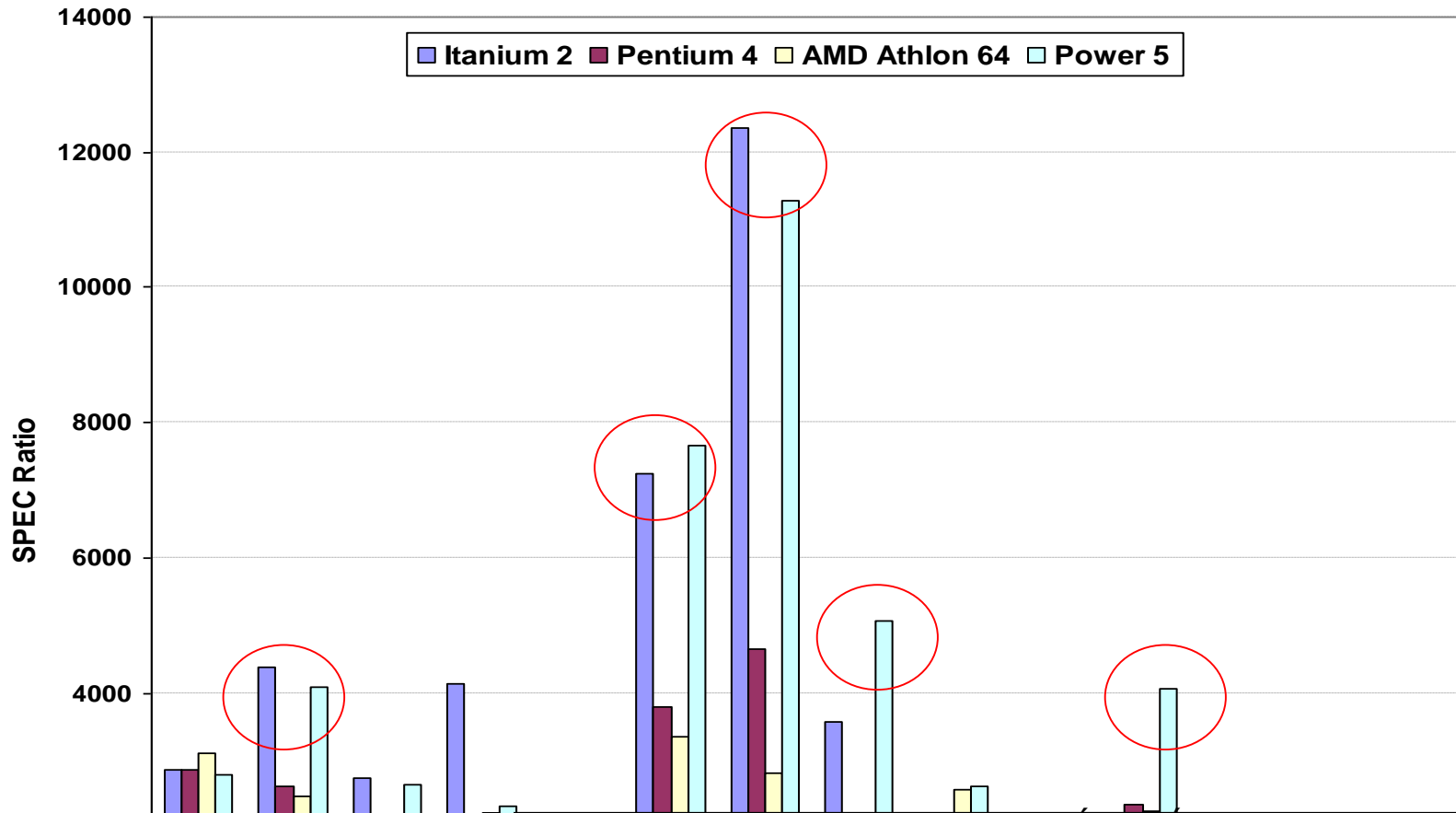
Cartagena99

CLASES PARTICULARES, TUTORIAS TECNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Rendimiento

❑ SPECfp2000



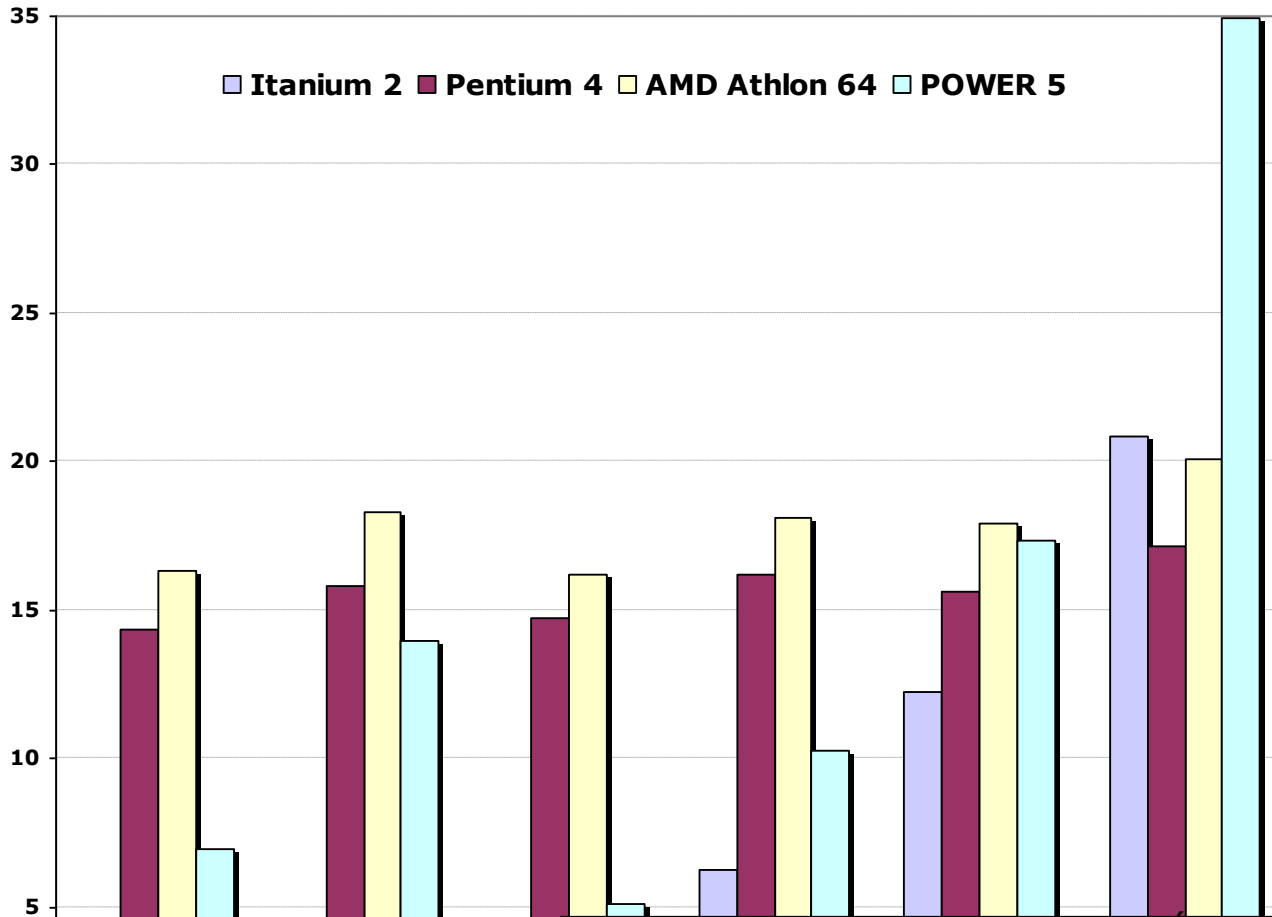
CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Cartagena99

Rendimiento

□ Rendimiento normalizado: Eficiencia



	I t a n i u m 2	P e n t i u m 4	A t h l o n 6 4	P o w e r 5
Rank	2	4	1	5
Int/Trans	4	2	1	3
FP/Trans	4	2	1	3
Int/area	4	2	1	3
FP/area	4	2	1	3
Int/Watt	4	3	1	2
FP/Watt	2	4	3	1

Cartagena99

CLASES PARTICULARES, TUTORIAS TECNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Transistors Transistors mm² mm² Watt Watt

Rendimiento Conclusiones

- ❑ No hay un claro ganador en todos los aspectos
- ❑ El AMD Athlon gana en SPECInt seguido por el Pentium4, Itanium 2, y Power5
- ❑ Itanium 2 y Power5, tienen similares rendimientos en SPECFP, dominan claramente al Athlon y Pentium 4
- ❑ Itanium 2 es el procesador menos eficiente en todas las medidas menos en SPECFP/Watt.
- ❑ Athlon y Pentium 4 usan bien los transistores y el área en términos de eficacia
- ❑ IBM Power5 es el mas eficaz en el uso de la energía sobre las SPECFP y algo menos sobre SPECINT

Cartagena99

CLASES PARTICULARES, TUTORIAS TECNICAS ONLINE
LLAMA O ENVIA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Conclusiones -Límites del ILP

- ❑ Doblar en ancho de emisión (issue rates) sobre los valores actuales 3-6 instrucciones por ciclo, a digamos 6 a 12 instrucciones requiere en el procesador
 - o de 3 a 4 accesos a cache de datos por ciclo,
 - o Predecir-resolver de 2 a 3 saltos por ciclo,
 - o Renombrar y acceder a mas de 20 registros por ciclo,
 - o Buscar de la cache de instrucciones de 12 a 24 instrucciones por ciclo.

- ❑ La complejidad de implementar estas capacidades implica al menos sacrificar la duración del ciclo e incrementa de forma muy importante el consumo.

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Conclusiones -Límites del ILP

- ❑ La mayoría de las técnicas que incrementan el rendimiento incrementan también el consumo.
- ❑ Una técnica es *eficiente en energía* si incrementa más el rendimiento que el consumo.
- ❑ Todas las técnicas de emisión múltiple son poco eficientes desde el punto de vista de la energía.

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70

Conclusiones -Límites del ILP

- ❑ En lugar de seguir explotando el ILP, los diseñadores se han focalizado sobre multiprocesadores en un chip (CMP, multicores,..)
- ❑ En el 2000, IBM abrió el campo con el 1º multiprocesador en un chip, el Power4, que contenía 2 procesadores Power3 y una cache L2 compartida. A partir de este punto todos los demás fabricantes han seguido el mismo camino. (Intel, AMD, Sun, Fujitsu,..).
- ❑ La explotación del paralelismo existente en las aplicaciones (a nivel de procesos y/o a nivel de datos) se plantea como la vía adecuada para obtener balances satisfactorios entre velocidad de procesamiento y consumo de energía.

Cartagena99

CLASES PARTICULARES, TUTORÍAS TÉCNICAS ONLINE
LLAMA O ENVÍA WHATSAPP: 689 45 44 70

ONLINE PRIVATE LESSONS FOR SCIENCE STUDENTS
CALL OR WHATSAPP:689 45 44 70