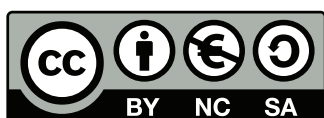


Arquitectura de Computadores

Paula Villanueva Núñez

Doble Grado de Informática y Matemáticas
Universidad de Granada



Este libro se distribuye bajo una licencia CC BY-NC-SA 4.0.

Eres libre de distribuir y adaptar el material siempre que reconozcas a los autores originales del documento, no lo utilices para fines comerciales y lo distribuyas bajo la misma licencia.

creativecommons.org/licenses/by-nc-sa/4.0/

Arquitectura de Computadores

Paula Villanueva Núñez

Doble Grado de Informática y Matemáticas

Universidad de Granada

Índice

1. Tema 1. Arquitecturas paralelas: clasificación y prestaciones	2
1.1. Lección 1. Clasificación del paralelismo implícito en una aplicación	2
1.1.1. 1.1 Objetivos	2
1.1.2. 1.2 Criterios de clasificaciones del paralelismo implícito en una aplicación.	2
1.1.3. 1.3 Dependencias de datos.	3
1.1.4. 1.4 Paralelismo implícito (nivel de detección), explícito y arquitecturas paralelas. . .	4
1.1.5. 1.5 Detección, utilización, implementación y extracción del paralelismo.	5
1.2. Lección 2. Clasificación de arquitecturas paralelas.	6
1.2.1. 2.1 Objetivos	6
1.2.2. 2.2 Computación paralela y computación distribuida.	7
1.2.3. 2.3 Clasificaciones de arquitecturas y sistemas paralelos.	7
1.2.4. 2.4 Nota histórica	21
1.3. Lección 3. Evaluación de prestaciones.	22
1.3.1. 3.1 Objetivos.	22
1.3.2. 3.2 Medidas usuales para evaluar prestaciones.	22
1.3.3. 3.3 Conjunto de programas de prueba (Benchmark).	26
1.3.4. 3.3.1 LINPACK.	27
1.3.5. 3.4 Ganancia en prestaciones.	28

1 Tema 1. Arquitecturas paralelas: clasificación y prestaciones

1.1 Lección 1. Clasificación del paralelismo implícito en una aplicación

1.1.1 1.1 Objetivos

- Clasificaciones del paralelismo implícito en una aplicación. Distinguir entre paralelismo de tareas y de datos.
- Distinguir entre dependencias RAW, WAW, WAR.
- Distinguir entre thread y proceso.
- Relacionar el paralelismo implícito en una aplicación con el nivel en el que se hace explícito para que se pueda utilizar (instrucción, thread, proceso) y con las arquitecturas paralelas que lo aprovechan.

1.1.2 1.2 Criterios de clasificaciones del paralelismo implícito en una aplicación.

- **Paralelismo funcional.**
 - **Nivel de funciones.** Las funciones llamadas en un programa se pueden ejecutar en paralelo, siempre que no haya entre ellas dependencias inevitables, como dependencias de datos verdaderas (lectura después de escritura).
 - **Nivel de bucle (bloques).** Se pueden ejecutar en paralelo las iteraciones de un bucle, siempre que se eliminen los problemas derivados de dependencias verdaderas. Para detectar dependencias habrá que analizar las entradas y las salidas de las iteraciones del bucle.
 - **Nivel de operaciones.** Las operaciones independientes se pueden ejecutar en paralelo. En los procesadores de propósito específico y en los de propósito general podemos encontrar instrucciones compuestas de varias operaciones que se aplican en secuencia al mismo flujo de datos de entrada. Se pueden usar instrucciones compuestas, que van a evitar las penalizaciones por dependencias verdaderas.
- **Paralelismo de datos** (*data parallelism* o *DLP-Data Level Par.*). Se encuentra implícito en las operaciones con estructuras de datos (vectores y matrices). Se puede extraer de la representación matemática de la aplicación. Las operaciones vectoriales y matriciales engloban operaciones que se pueden realizar en paralelo. Por lo que el paralelismo de datos está relacionado con el paralelismo a nivel de bucle.
- **Paralelismo de tareas** (*task parallelism* o *TLP-Task Level Par.*). Se encuentra extrayendo la estructura lógica de funciones de una aplicación. Los bloques son funciones y se puede encontrar paralelismo entre las funciones.
- **Granularidad.** El grano más pequeño (*grano fino*) se asocia al paralelismo entre operaciones o instrucciones, el *grano medio* se asocia a los bloques funcionales lógicos y el *grano grueso* se asocia al paralelismo entre programas.

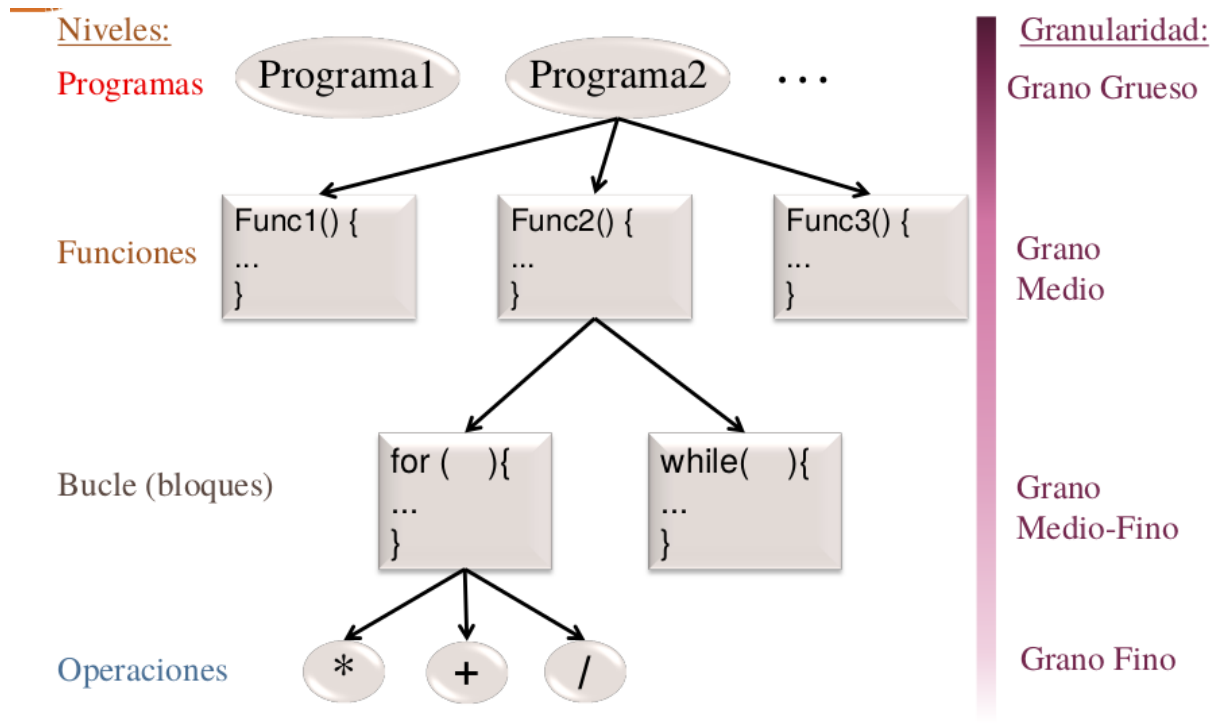


Figura 1:

1.1.3 1.3 Dependencias de datos.

Para que un bloque de código B_2 presente dependencia de datos con respecto a B_1 , deben hacer referencia a una misma posición de memoria M (variable) y B_1 aparece en la secuencia de código antes que B_2 .

Tipos de dependencias de datos (de B_2 respecto a B_1):

- **RAW** (*Read After Write*) o dependencia verdadera.
- **WAW** (*Write After Write*) o dependencia de salida.
- **WAR** (*Write After Read*) o antidependencia.

```

1 ...
2 a = b + c
3 ... //código que no usa a
4 d = a + c
5 ...

```

```

1 ...
2 a = b + c
3 ... //se lee a
4 a = d + e
5 ... //se lee a

```

```

1 ...
2 b = a + 1

```

```

3 ...
4 a = d + e
5 ... //se lee a

```

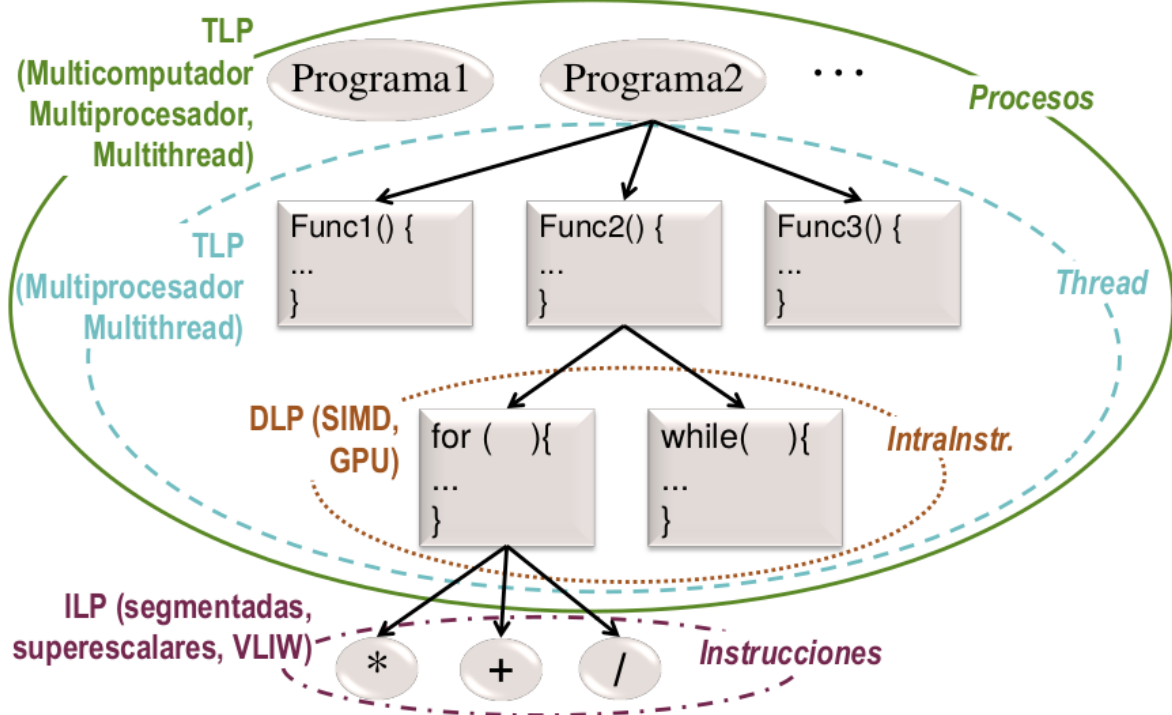
1.1.4 1.4 Paralelismo implícito (nivel de detección), explícito y arquitecturas paralelas.

El paralelismo entre **programas** se utiliza a nivel de procesos. Cuando se ejecuta un programa, se crea el proceso asociado al programa.

El paralelismo entre **funciones** se puede extraer para utilizarlo a nivel de procesos o de hebras.

El paralelismo dentro de un **bucle** se puede extraer a nivel de procesos o de hebras. Se puede aumentar la granularidad asociando un mayor número de iteraciones del ciclo a cada unidad a ejecutar en paralelo. Se puede hacer explícito dentro de una instrucción vectorial para que sea aprovechado por arquitecturas SIMD o vectoriales.

El paralelismo entre **operaciones** se puede aprovechar en arquitecturas con paralelismo a nivel de instrucción (ILP) ejecutando en paralelo las instrucciones asociadas a estas operaciones independientes.



1.4.1 Nivel de paralelismo explícito.

1.4.1.1 Unidades en ejecución en un computador.

- **Instrucciones.** La unidad de control de un core o procesador gestiona la ejecución de instrucciones por la unidad de procesamiento.
- **Thread o light process.** Es la menor unidad de ejecución que gestiona el SO. Menor secuencia de instrucciones que se pueden ejecutar en paralelo o concurrentemente.

- **Proceso o process.** Mayor unidad de ejecución que gestiona el SO. Un proceso consta de uno o varios thread.

1.4.1.2 Threads versus procesos.

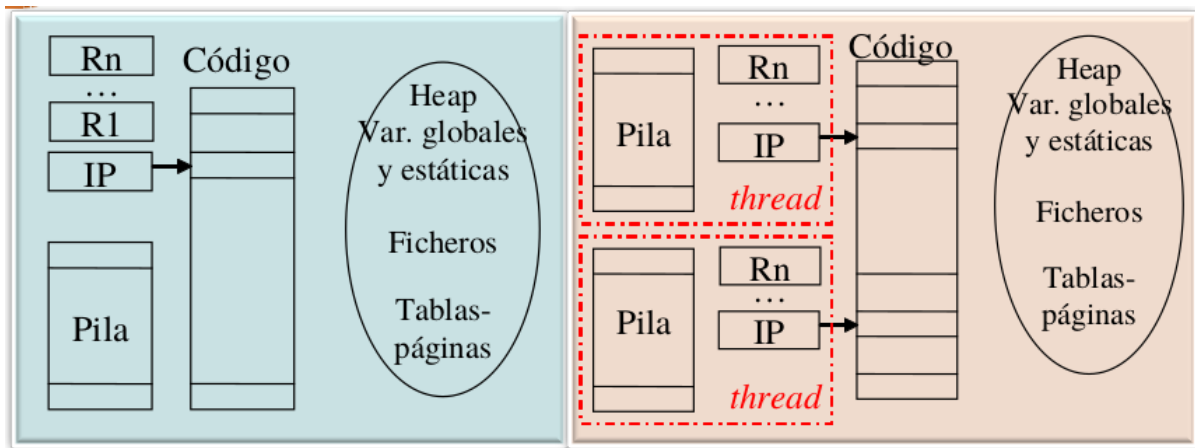
El hardware gestiona la ejecución de las instrucciones. A nivel superior, el SO se encarga de gestionar la ejecución de unidades de mayor granularidad, procesos y hebras. Cada proceso en ejecución tiene su propia asignación de memoria. Los SO **multihebra** permiten que un proceso se componga de una o varias hebras (hilos). Una **hebra** tiene su propia pila y contenido de registros, entre ellos el puntero de pila y el IP (Puntero de Instrucciones) que almacena la dirección de la siguiente instrucción a ejecutar de la hebra, pero comparte el código, las variables globales y otros recursos con las hebras del mismo proceso. Por lo que las hebras se pueden crear y destruir en menor tiempo que los procesos, y la comunicación (se usa la memoria que comparten), sincronización y conmutación entre hebras de un proceso es más rápida que entre procesos. Luego las hebras tienen menor granularidad que los procesos.

Un **proceso** comprende el código del programa y todo lo que hace falta para su ejecución:

- Datos en pila, segmentos (variables globales y estáticas) y en heap (BP1).
- Contenido de los registros.
- Tabla de páginas.
- Tabla de ficheros abiertos.

Para comunicar procesos hay que usar llamadas al SO.

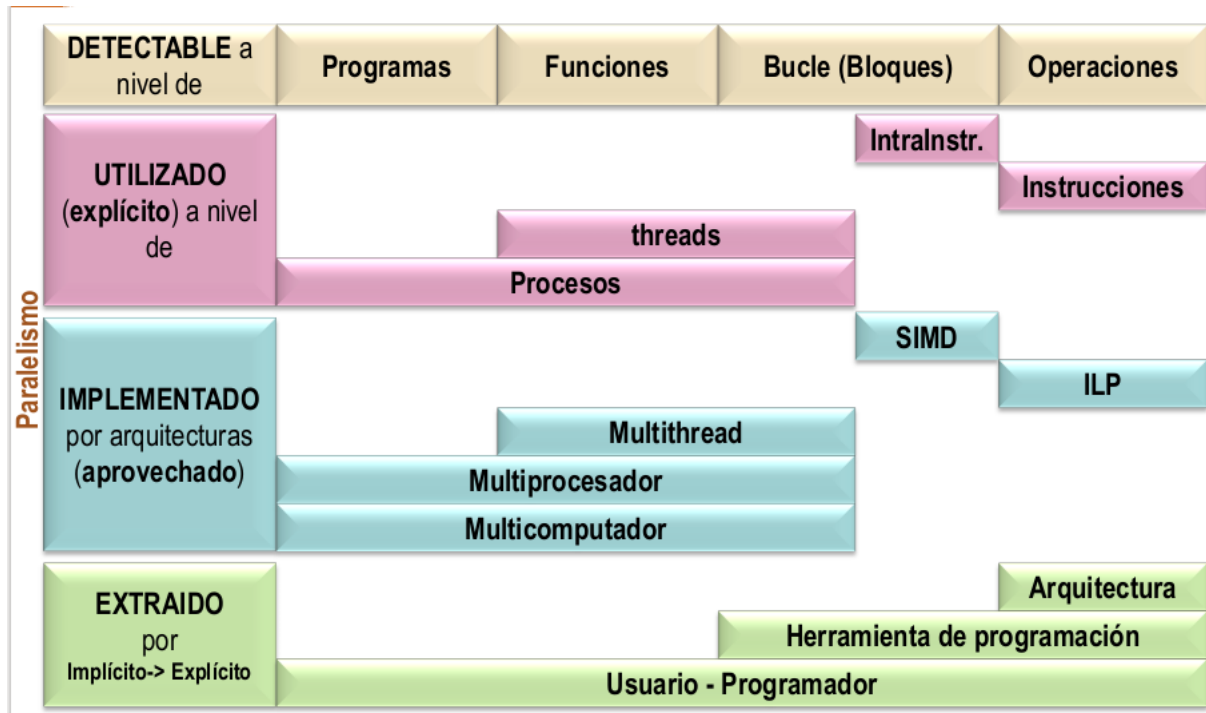
El paralelismo implícito en el código de una aplicación se puede hacer explícito a nivel de instrucciones, de hebras o de procesos.



1.1.5 1.5 Detección, utilización, implementación y extracción del paralelismo.

En los procesadores ILP superescalares o segmentados la arquitectura extrae paralelismo. Para ello, eliminan dependencias de datos falsas entre instrucciones y evitan problemas debidos a dependencias de datos, de control y de recursos. La arquitectura extrae paralelismo implícito en las entradas en tiempo de ejecución (dinámicamente). El grado de paralelismo de las instrucciones aprovechado se puede incrementar con ayuda del compilador y del programador. Podemos definir el grado de paralelismo de

un conjunto de entradas a un sistema como el máximo número de entradas del conjunto que se puede ejecutar en paralelo. Para los procesadores las entradas son instrucciones. En las arquitecturas ILP VLIW el paralelismo que se va a aprovechar está ya explícito en las entradas. Las instrucciones que se van a ejecutar en paralelo se captan juntas de memoria. El análisis de dependencias entre instrucciones en este caso es estático.



Hay compiladores que extraen el paralelismo de datos implícito a nivel de bucle. Algunos compiladores lo hacen explícito a nivel de hebra, y otros dentro de una instrucción para que se pueda aprovechar en arquitecturas SIMD o vectoriales. El usuario, como programador, puede extraer el paralelismo implícito en un bucle o entre funciones definiendo hebras y/o procesos. La distribución de las tareas independientes entre hebras o entre procesos dependerán de

- la granularidad de las unidades de código independientes,
- la posibilidad que ofrezca la herramienta para programación paralela disponible de definir hebras o procesos,
- la arquitectura disponible para aprovechar el paralelismo,
- el SO disponible.

Por último, los usuarios del sistema al ejecutar programas están creando procesos que se pueden ejecutar en el sistema concurrentemente o en paralelo.

1.2 Lección 2. Clasificación de arquitecturas paralelas.

1.2.1 2.1 Objetivos

- Distinguir entre procesamiento o computación paralela y distribuida.

- Clasificar los computadores según segmento del mercado.
- Distinguir entre las diferentes clases de arquitecturas de la clasificación de Flynn.
- Diferenciar un multiprocesador de un multicomputador.
- Distinguir entre NUMA y SMP.
- Distinguir entre arquitecturas DLP, ILP, TLP.
- Distinguir entre arquitecturas TLP con una instancia de SO y TLP con varias instancias de SO.

1.2.2 2.2 Computación paralela y computación distribuida.

- **Computación paralela.** Estudia los aspectos hardware y software relacionados con el desarrollo y ejecución de aplicaciones en un sistema de cómputo compuesto por múltiples cores/procesadores/computadores que es visto externamente como una unidad autónoma (multicores, multiprocesadores, multicomputadores, cluster).
- **Computación distribuida.** Estudia los aspectos hardware y software relacionados con el desarrollo y ejecución de aplicaciones en un sistema distribuido; es decir, en una colección de recursos autónomos (PC, servidores -de datos, software, . . . -, supercomputadores. . .) situados en distintas localizaciones físicas.
 - **Computación distribuida baja escala.** Estudia los aspectos relacionados con el desarrollo y ejecución de aplicaciones en una colección de recursos autónomos de un dominio administrativo situados en distintas localizaciones físicas conectados a través de infraestructura de red local.
 - **Computación distribuida gran escala.**
 - **Computación grid.** Estudia los aspectos relacionados con el desarrollo y ejecución de aplicaciones en una colección de recursos autónomos de múltiples dominios administrativos geográficamente distribuidos conectados con infraestructura de telecomunicaciones.
 - **Computación cloud.** Comprende los aspectos relacionados con el desarrollo y ejecución de aplicaciones en un sistema cloud. El sistema cloud ofrece servicios de infraestructura, plataforma y/o software, por los que se paga cuando se necesitan (pay-per-use) y a los que se accede típicamente a través de una interfaz (web) de auto-servicio. El sistema cloud consta de recursos virtuales que son una abstracción de los recursos físicos, parecen ilimitados en número y capacidad y son reclutados/liberados de forma inmediata sin interacción con el proveedor, soportan el acceso de múltiples clientes (multitenant) y están conectados con métodos estándar independientes de la plataforma de acceso.

1.2.3 2.3 Clasificaciones de arquitecturas y sistemas paralelos.

2.3.1 Criterios de clasificación de computadores

- Comercial. Segmento del mercado: embebidos, servidores gama baja. . .

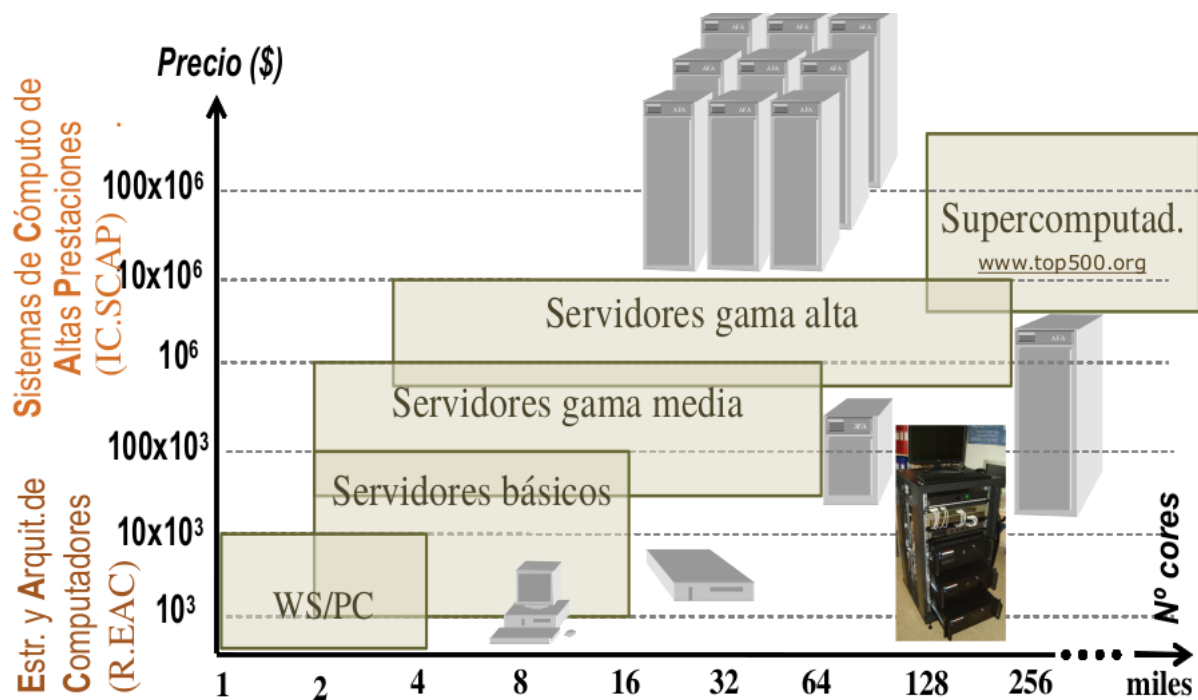


- Educación, investigación (también usados por fabricantes y vendedores):
 - Flujos de control y flujos de datos: clasificación de Flynn.
 - Sistemas de memoria.
 - Flujos de control (propuesta de clasificación de arquitecturas con múltiples flujos de control).
 - Nivel del paralelismo aprovechado (propuesta de clasificación).

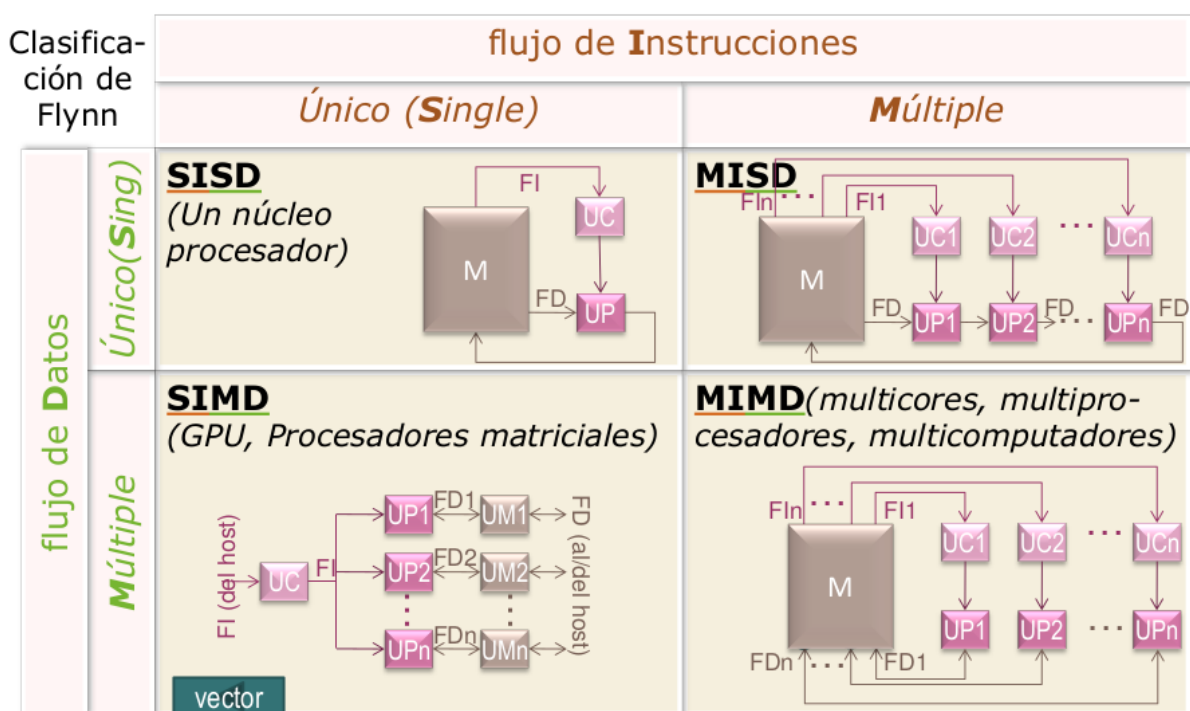
2.3.2 Clasificación de computadores según segmento

- **Externo** (*desktop, laptop, server, cluster...*) - R.EAC, IC.SCAP. Para todo tipo de aplicaciones:
 - Oficina, entretenimiento...
 - Procesamiento de transacciones o OLTP, sistemas de soporte de decisiones o DSS, e-commerce...
 - Científicas (medicina, biología, predicción del tiempo...) y animación (películas animadas, efectos especiales...).
- **Empotrado** (oculto) - IC.SCAE. Aplicaciones de propósito específico (videojuegos, teléfonos, coches, electrodomésticos...). Las restricciones típicas son: consumo de potencia, precio, tamaño reducido, tiempo real...

2.3.3 Clasificación de computadores externos según segmento del mercado



2.3.4 Clasificación de Flynn de arquitecturas (flujo instrucciones / flujo de datos)

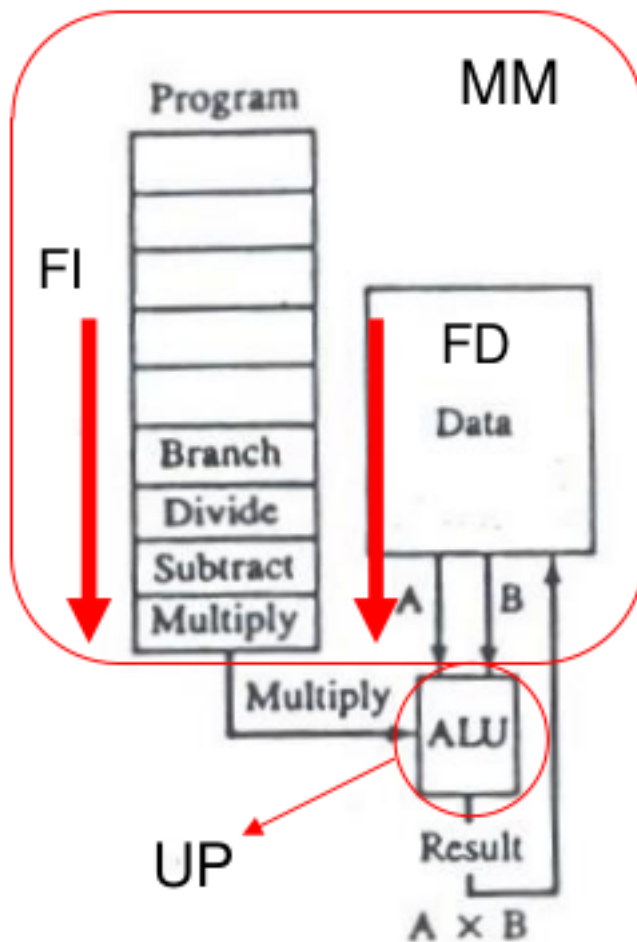


- **Computadores SISD.** Un único flujo de instrucciones (SI) procesa operandos y genera resultados, definiendo un único flujo de datos (SD).

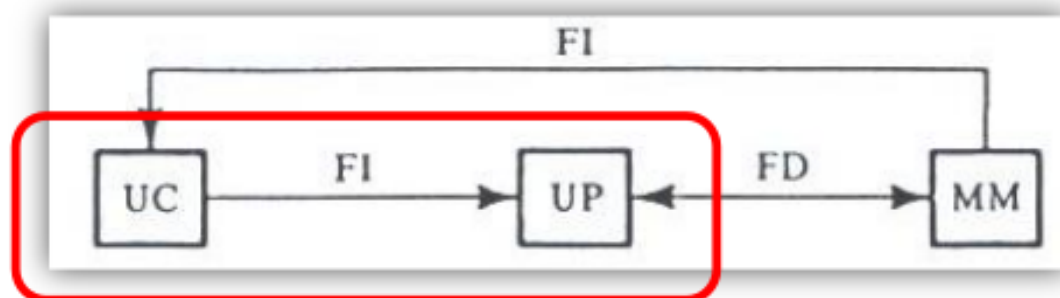
Corresponde a los computadores uni-procesador, ya que existe una única unidad de control (UC) que recibe las instrucciones de memoria, las decodifica y genera los códigos que definen la operación correspondiente a cada instrucción que debe realizar la unidad de procesamiento

(UP) de datos.

Descripción Funcional



Descripción Estructural

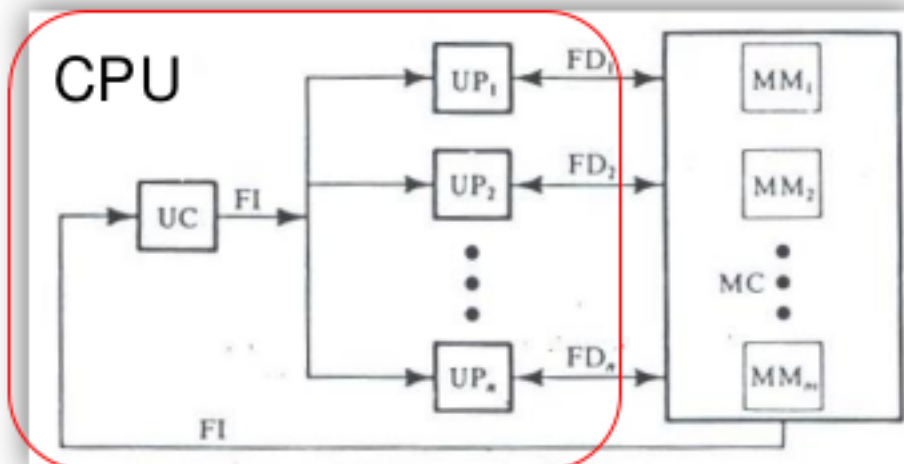


CPU, núcleo, procesador

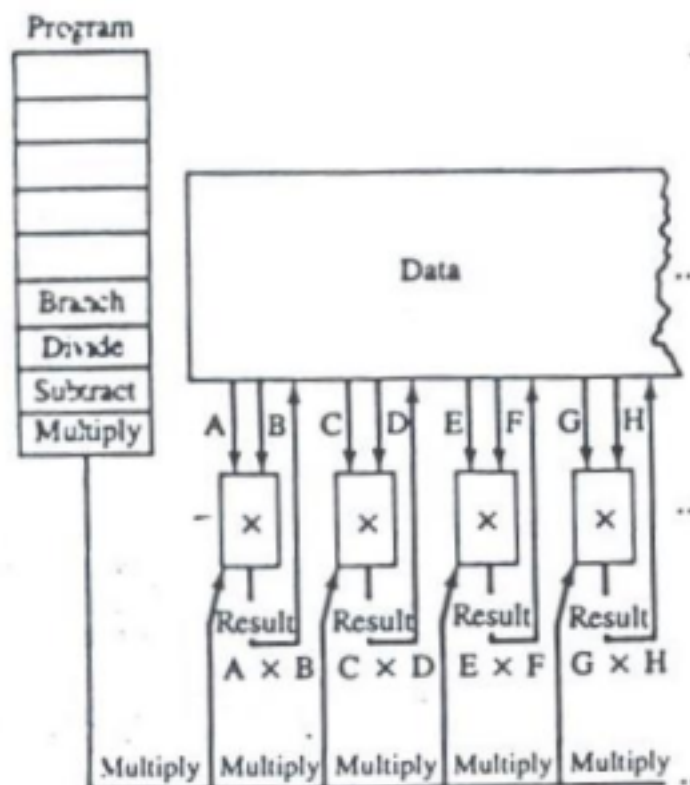
- **Computadores SIMD.** Un único flujo de instrucciones (SI) procesa operandos y genera resultados, definiendo varios flujos de datos (MD), dado que cada instrucción codifica realmente varias operaciones iguales, cada una actuando sobre operadores distintos.

Los códigos que genera la única unidad de control a partir de cada instrucción actúan sobre varias unidades de procesamiento distintas (UP_i). Por lo que se pueden realizar varias operaciones similares simultáneas con operandos distintos. Aprovechan paralelismo de datos.

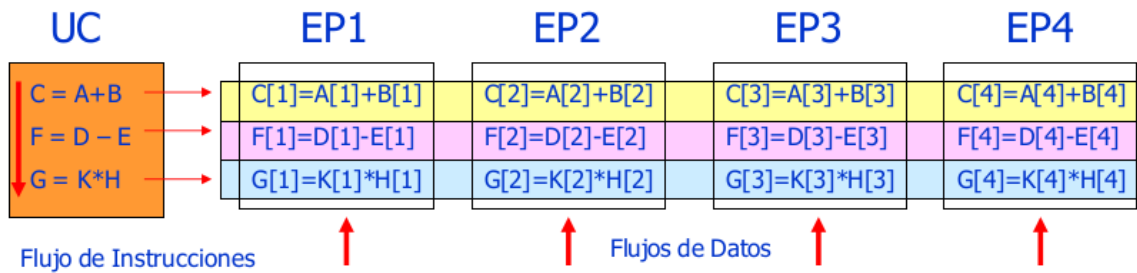
Descripción Estructural



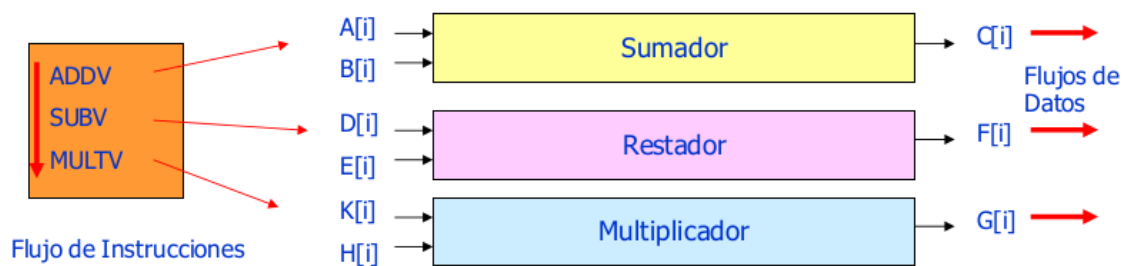
Descripción Funcional



Procesador Matricial

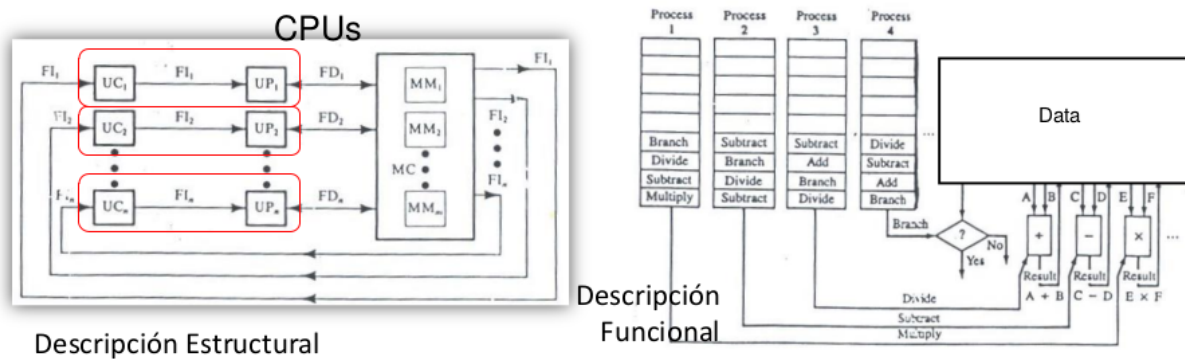


Procesador Vectorial

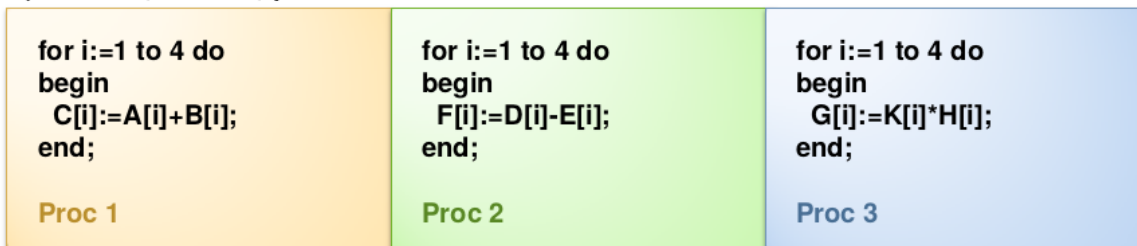


- **Computadores MIMD.** El computador ejecuta varias secuencias o flujos distintos de instrucciones (MI) y cada uno de ellos procesa operandos y genera resultados definiendo un único flujo de instrucciones, de forma que existen varios flujos de datos (MD) uno por cada flujo de instrucciones.

Corresponde con multinúcleos, multiprocesadores y multicomputadores. Puede aprovechar paralelismo funcional. Existen varias unidades de control que decodifican las instrucciones correspondientes a distintos programas. Cada uno de estos programas procesa conjuntos de datos diferentes, que definen distintos flujos de datos.



Corresponde con Multinúcleos, Multiprocesadores y Multicomputadores: Puede aprovechar, además, **paralelismo funcional**



- **Computadores MISD.** Se ejecutan varios flujos distintos de instrucciones (MI) aunque todos actúan sobre el mismo flujo de datos (SD).

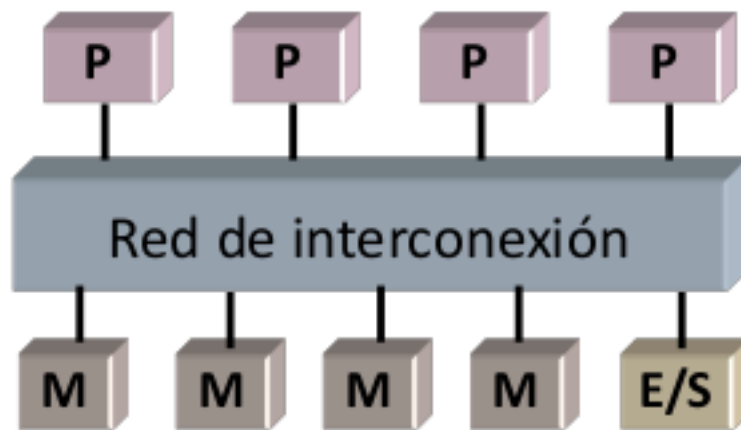
No existen computadores que funcionen según este modelo. Se puede simular en un código este modelo para aplicaciones que procesan una secuencia o flujo de datos.

2.3.5 Sistemas de memoria

2.3.5.1 Clasificación de computadores paralelos MIMD según el sistema de memoria

Los sistemas multiprocesadores se han clasificado atendiendo a la organización del sistema de memoria:

- **Sistemas con memoria compartida (SM) o multiprocesadores.** Son sistemas en los que todos los procesadores comparten el mismo espacio de direcciones. El programador no necesita conocer dónde están almacenados los datos.
- **Sistemas con memoria distribuida (DM) o multicomputadores.** Son sistemas en los que cada procesador tiene su propio espacio de direcciones particular. El programador necesita conocer dónde están almacenados los datos.



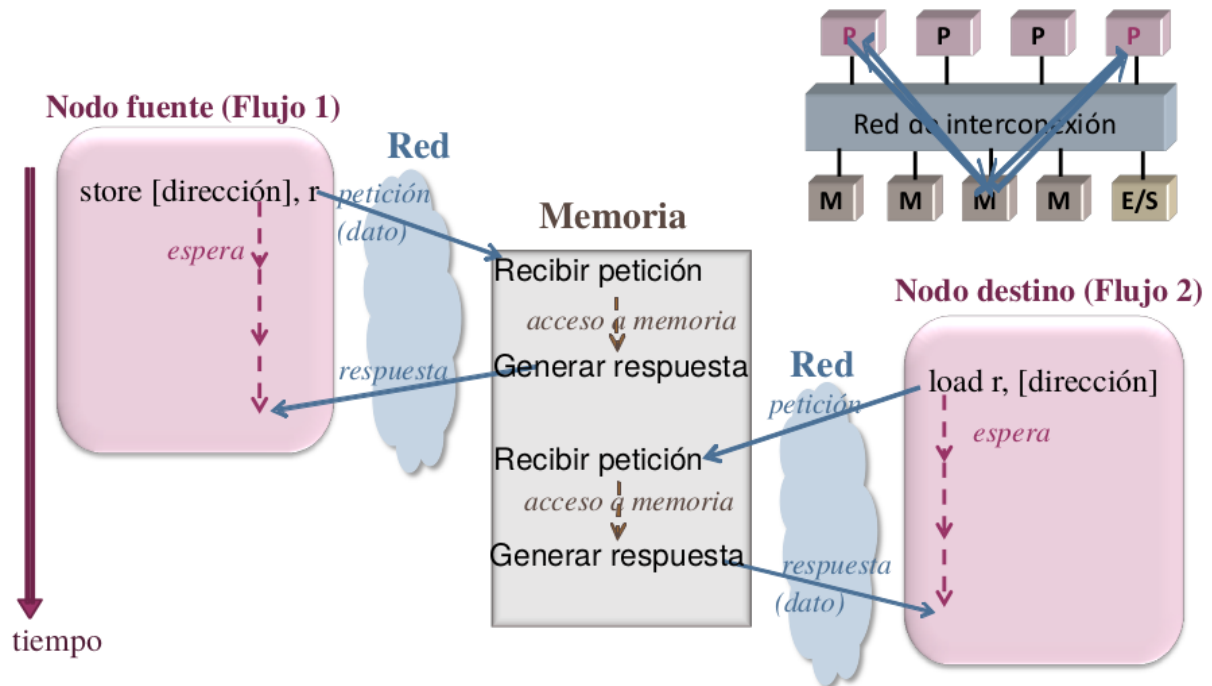
2.3.5.2 Comparativa SMP (Symmetric MultiProcessor) y multicomputadores.

- **Multiprocesador con memoria centralizada (SMP).** Es un multiprocesador en el que el tiempo de acceso de los procesadores a memoria es igual sea cual sea la posición de memoria a la que acceden, es una estructura simétrica.
 - Mayor latencia.
 - Poco escalable.
 - La comunicación es implícita mediante variables compartidas.
 - Los datos no están duplicados en memoria principal.
 - Necesita implementar primitivas de sincronización.
 - La distribución de código y datos entre procesadores no es necesaria.
 - La programación es más sencilla.
- **Multicomputador.**
 - Menor latencia.
 - Más escalable.
 - La comunicación es explícita mediante software para paso de mensajes (*send/receive*).
 - Los datos están duplicados en memoria principal y se copian datos entre módulos de memoria de diferentes procesadores.
 - La sincronización se hace mediante software de comunicación.
 - La distribución de código y datos entre procesadores es necesaria y se necesitan herramientas de programación más sofisticadas.
 - La programación es más difícil.

2.3.5.3 Comunicación uno-a-uno en un multiprocesador.

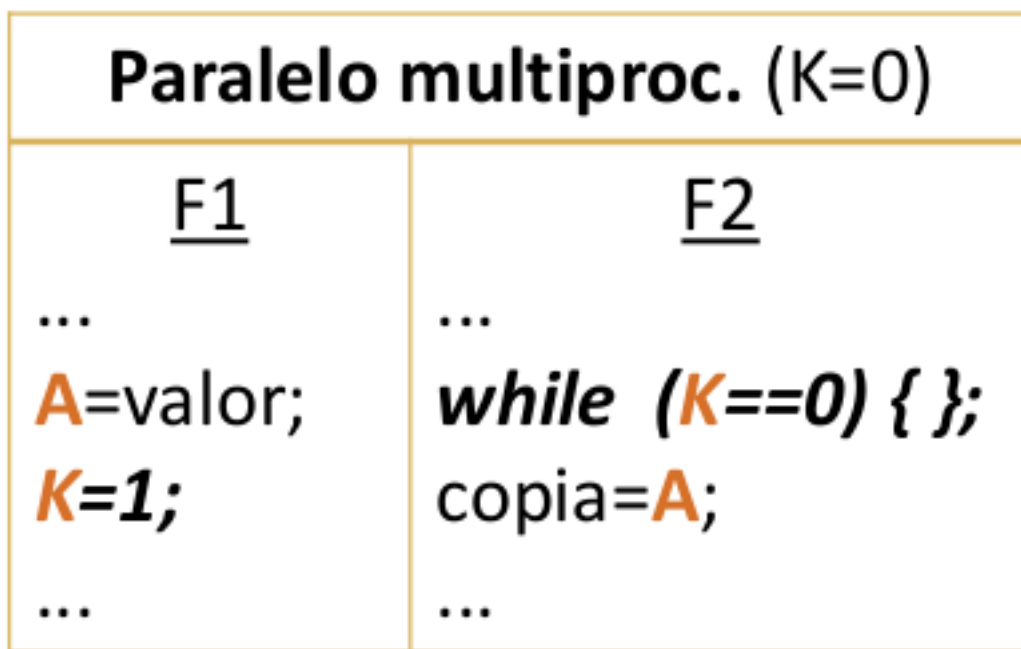
Los diferentes procesadores que ejecutan una aplicación pueden requerir sincronizarse en algún momento. Por ejemplo, si el procesador A utiliza un dato que produce el procesador B, A deberá esperar a que B lo genere. En la siguiente imagen vemos la transferencia de datos en un multiprocesador. El proceso que ejecuta la instrucción de carga espera hasta recibir el contenido de la dirección. El proceso que ejecuta la instrucción de almacenamiento puede esperar a que termine para garantizar que se mantiene un orden

en los accesos a memoria. Obsérvese que para que la transferencia de datos se realice de forma efectiva habría que sincronizar los procesos fuente y destino.



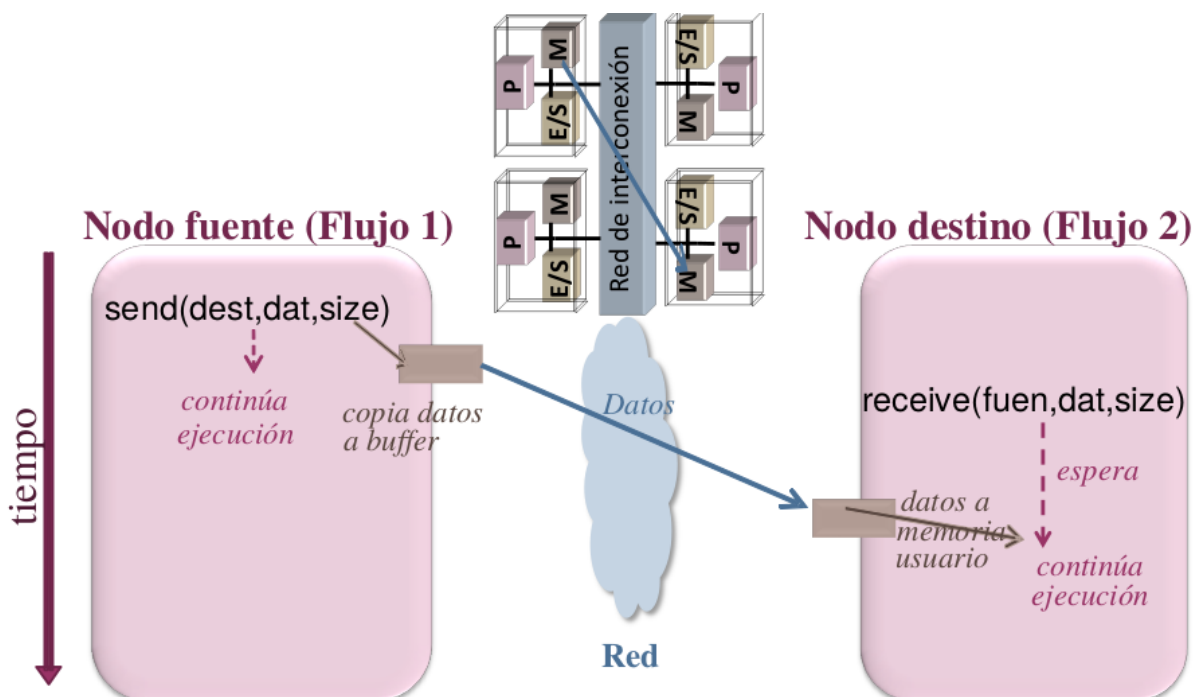
Secuencial	Paralelo	
...	F1	F2
A=valor;
...	A=valor;	copia=A;
copia=A;
...		

F1 es el flujo de control productor del dato (envía el dato)
F2 es el flujo de control consumidor del dato (recibe el dato)



Se debe garantizar que el flujo de control **consumidor** del dato lea la variable compartida (A) cuando el **productor** haya escrito en la variable el dato.

En multicomputadores se aprovechan los mecanismos de comunicación para implementar sincronización. Con una función de recepción bloqueante, es decir, que deja al proceso que la ejecuta detenido hasta que se reciba el dato, se puede implementar sincronización. En la siguiente figura podemos ver la transferencia asíncrona (con función *receive* bloqueante) de datos en un multicomputador. EN transferencia asíncrona se requiere almacenamiento intermedio para evitar esperas. El proceso fuente continúa la ejecución en cuanto los datos se copien en un *buffer*. El destino espera en el *receive* bloqueante a que lleguen los datos, en caso de que estos no hayan llegado aún.



Secuencial	Paralelo	
...	<u>F1</u>	<u>F2</u>
A=valor;
...	A=valor;	copia=A;
copia=A;
...		

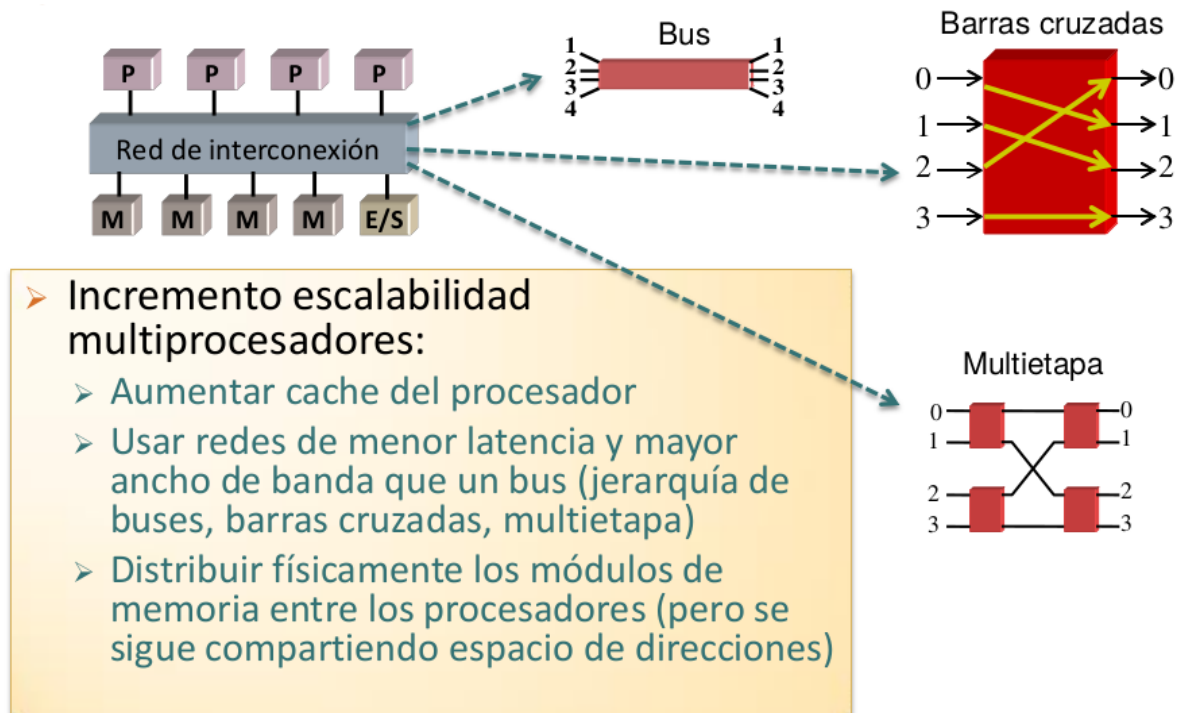
F1 es el flujo de control productor del dato (*envía el dato*)
F2 es el flujo de control consumidor del dato (*recibe el dato*)

Paralelo multicomputador (size = 4 byte)	
<u>F1</u>	<u>F2</u>
...	...
send(F2, valor, 4);	receive(F1,copia,4);
...	...

2.3.5.4 Incremento de escalabilidad en multiprocesadores y red de interconexión.

Como los SMP tienen escasa escalabilidad, se ha intentado incrementar la escalabilidad en multiprocesadores:

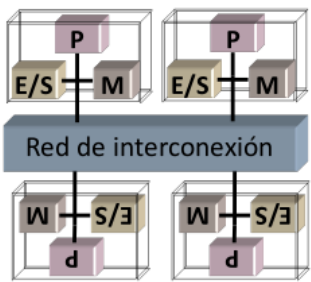
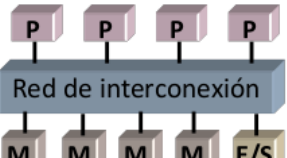
- Aumentar caché del procesador.
- Usar redes de menor latencia y mayor ancho de banda que un bus (jerarquía de buses, barras cruzadas, multietapa).
- Distribuir físicamente los módulos de memoria entre los procesadores (pero se sigue compartiendo espacio de direcciones).



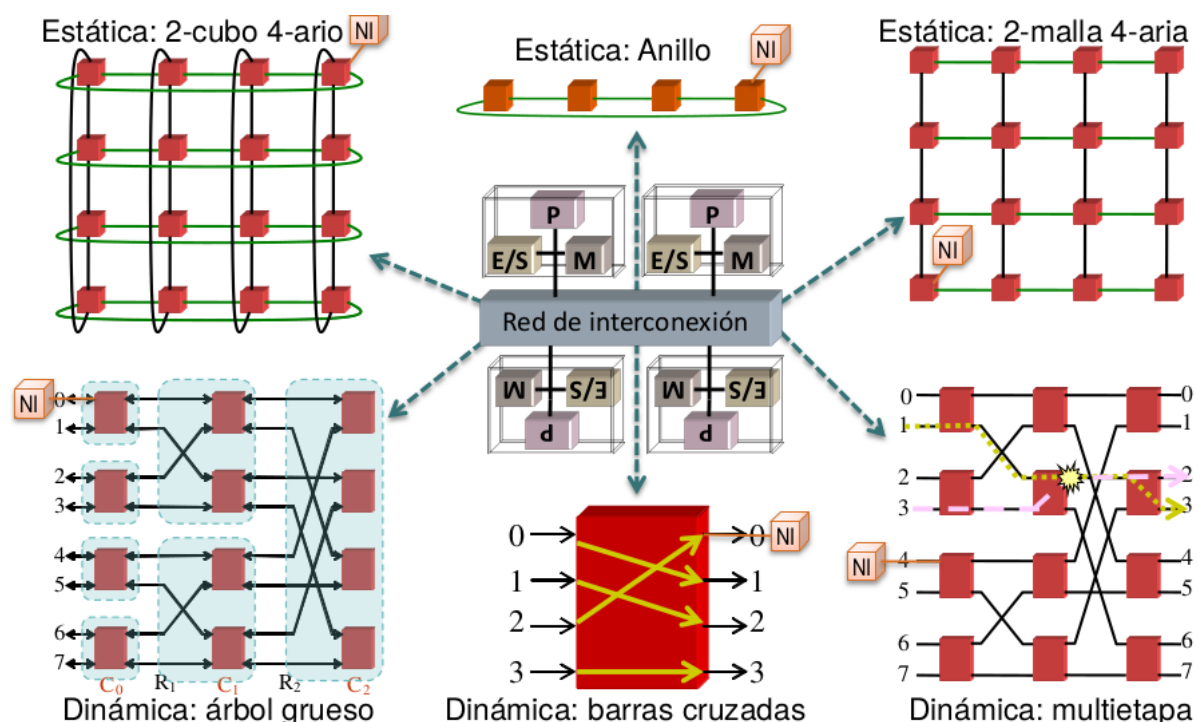
2.3.5.5 Clasificación completa de computadores según el sistema de memoria.

■ Multiprocesadores.

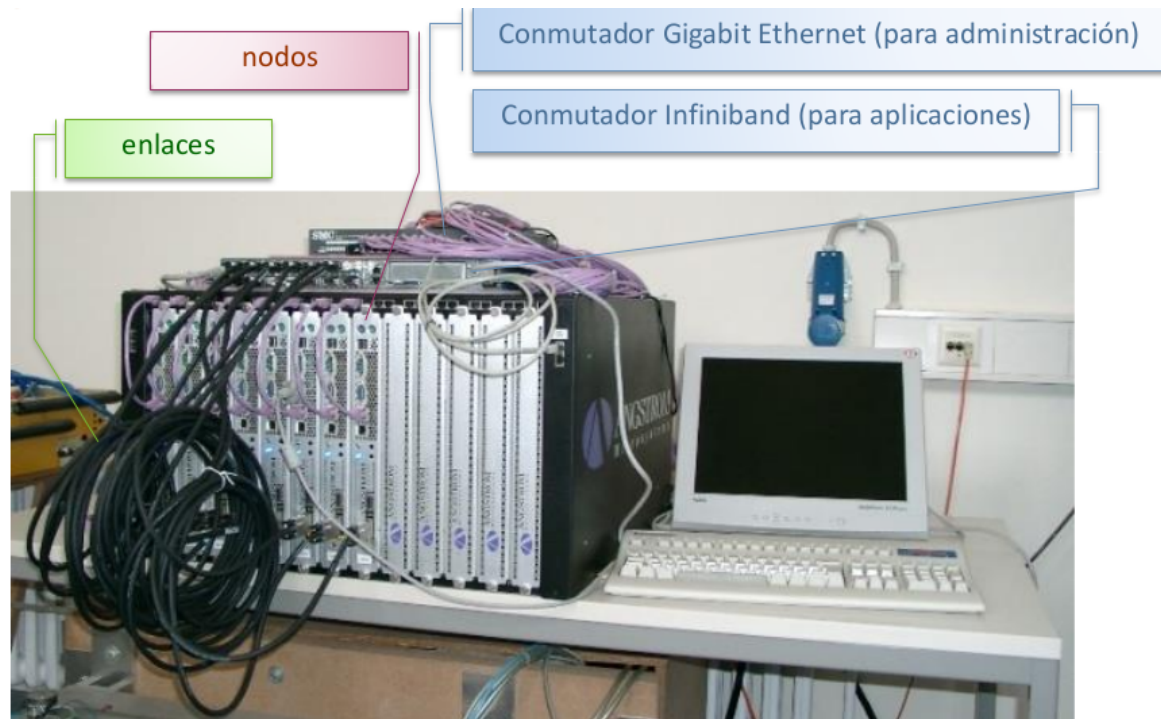
- **UMA** (*Uniform Memory Access*).
 - **SMP**.
- **NUMA** (*Non-Uniform Memory Access*).
 - **NUMA**. Arquitecturas con acceso a memoria no uniforme sin coherencia de caché entre nodos. No incorporan hardware para evitar problemas por incoherencias entre cachés de distintos nodos. Esto hace que los datos modificables compartidos no se puedan trasladar a caché de nodos remotos; hay que acceder a ellos individualmente a través de la red. Se puede hacer más tolerable la latencia utilizando precaptación (*prefetching*) de memoria y procesamiento multihebra.
 - **CC-NUMA**. Arquitecturas con acceso a memoria no uniforme y con caché coherente. Tienen hardware para mantener coherencia entre cachés de distintos nodos, que se encarga de las transferencias de datos compartidos entre nodos. El hardware para mantenimiento de coherencia supone un coste añadido e introduce un retardo que hace que estos sistemas escalen en menor grado que un NUMA.
 - **COMA**. Arquitecturas con acceso a memoria solo caché. La memoria local de los procesadores se gestiona como caché. El sistema de mantenimiento se encarga de llevar dinámicamente el código y los datos a los nodos donde se necesitan.

Multi-computadores Memoria no compartida	NORMA <i>No Remote Memory Access</i>	<i>ej. cluster, red de computadores</i>	Memoria físicamente distribuida	+	+
Multi-procesadores Memoria compartida Un único espacio de direcciones	NUMA <i>Non-Uniform Memory Access</i>	NUMA		Escalabilidad	-
		CC-NUMA			
		COMA			
	UMA <i>Uniform Memory Access</i>	SMP <i>Symmetric MultiProcessor</i>	Memoria físicamente centralizada	-	-
					

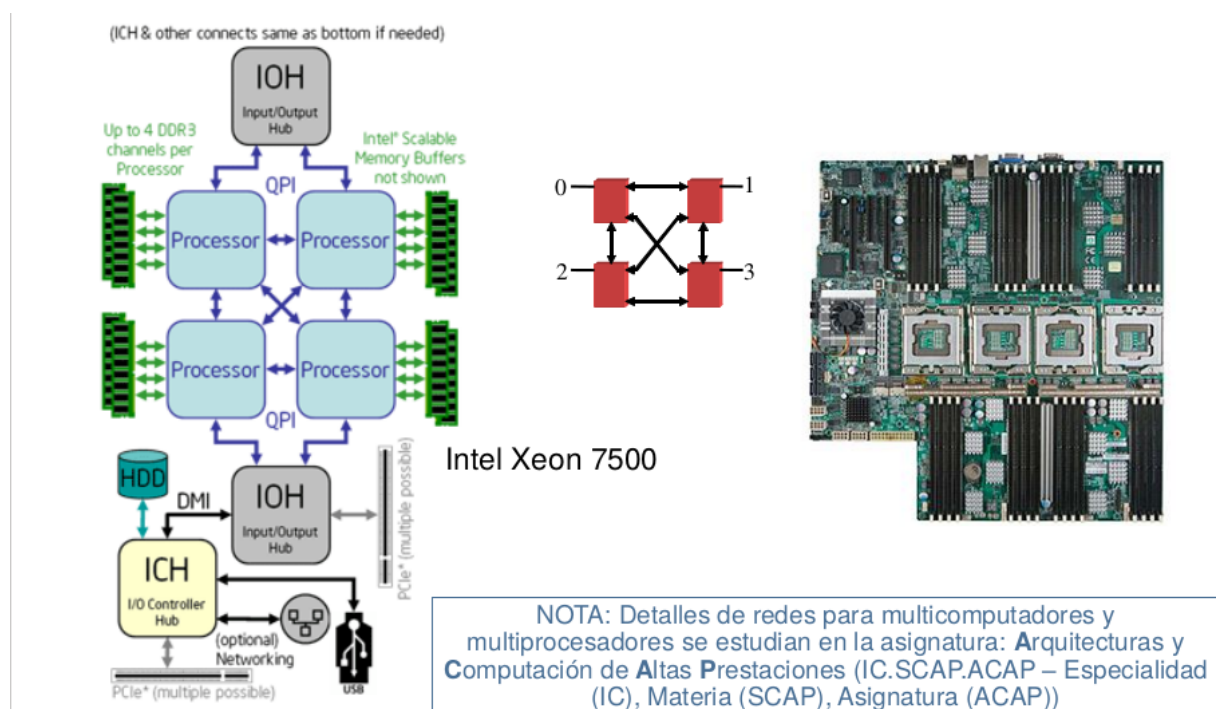
2.3.5.6 Red en sistemas con memoria físicamente distribuida (NI: Network Interface).



Ejemplo: Red (con conmutador o switch) de barras cruzadas.



Ejemplo: Placa CC-NUMA con red estática



2.3.6 Flujos de control (propuesta de clasificación de arquitecturas con múltiples flujos de control (o threads o flujos de instrucciones))

- **TLP** (*Thread Level Parallelism*). Ej. múltiples flujos de control concurrentemente o en paralelo.
 - **Implícito**. Flujos de control creados y gestionados por la arquitectura.
 - **Explícito**. Flujos de control creados y gestionados por el SO.

- **Con una instancia SO.** Multiprocesadores, multicores, cores multithread. . .
- **Con múltiples instancias SO.** Multicomputadores.

2.3.7 Nivel del paralelismo aprovechado (propuesta de clasificación)

- **Arquitecturas con DLP** (*Data Level Parallelism*). Ejecutan las operaciones de una instrucción concurrentemente o en paralelo: unidades funcionales vectoriales o SIMD.
- **Arquitecturas con ILP** (*Instruction Level Parallelism*). Ejecutan múltiples instrucciones concurrentemente o en paralelo: cores escalares segmentados, superescalares o VLIW/EPIC.
- **Arquitecturas con TLP** (*Thread Level Parallelism*) explícito y una instancia de SO. Ejecutan múltiples flujos de control concurrentemente o en paralelo.
 - **Cores** que modifican la arquitectura escalar segmentada, superescalar o VLIW/EPIC para ejecutar threads concurrentemente o en paralelo.
 - **Multiprocesadores:** ejecutan threads en paralelo en un computador con múltiples cores (incluye multicores).
- **Arquitecturas con TLP explícito y múltiples instancias SO.** Ejecutan múltiples flujos de control en paralelo.
 - **Multicomputadores:** ejecutan threads en paralelo en un sistema con múltiples computadores.

1.2.4 2.4 Nota histórica

- **DLP** (*Data Level Parallelism*). Unidades funcionales (o de ejecución) SIMD (o multimedia).
- **ILP** (*Instruction Level Parallelism*).
 - Procesadores/cores segmentados.
 - Procesadores con múltiples unidades funcionales.
 - Procesadores/cores superescalares
 - Procesadores/cores VLIW
- **TLP** (*Thread Level Parallelism*).
 - TLP explícito con una instancia de SO.
 - Multithread grano fino (FGMT).
 - Multithread grano grueso (CGMT).
 - Multithread simultánea (SMT).
 - Multiprocesadores en un chip (CMP) o multicores.
 - Multiprocesadores.
 - TPL explícito con múltiples instancias del SO (multicomputadores): IC.SCAP.

1.3 Lección 3. Evaluación de prestaciones.

1.3.1 3.1 Objetivos.

- Distinguir entre tiempo de CPU (sistema y usuario) de unix y tiempo de respuesta.
- Distinguir entre productividad y tiempo de respuesta.
- Obtener, de forma aproximada mediante cálculos, el tiempo de CPU, GFLOPS y los MIPS del código ejecutado en un núcleo de procesamiento.
- Explicar el concepto de ganancia en prestaciones.
- Aplicar la ley de Amdahl.

1.3.2 3.2 Medidas usuales para evaluar prestaciones.

3.2.1 Tiempo de respuesta.

- Real (*wall-clock time, elapsed time, real time*).
- *CPU time = user + sys* (no incluye todo el tiempo).
- Con un flujo de control.
 - $\text{elapsed} \geq \text{CPU time}$.

```
1 time ./program.exe
2 elapsed 5.4
3 user 3.2
4 sys 1.0
```

- Con múltiples flujos de control
 - $\text{elapsed} < \text{CPU time}$, $\text{elapsed} \geq \text{CPU time} / n^\circ \text{ flujos control}$.

En el programa, *user 3.2* significa el tiempo de CPU de usuario (tiempo de ejecución en espacio de usuario). *sys 1.0* significa el tiempo de CPU de sistema (tiempo en el nivel del kernel del SO). Además, hay otro tiempo asociado a las esperas debidas a I/O o asociados a la ejecución de otros programas.

Comando *time* en Unix: *3.2u + 1.0s* es el 78 % del tiempo transcurrido (5.4).

Alternativas para obtener tiempos:

Función	Fuente	Tipo	Resolución aprox. (microsecs)
<i>time</i>	SO (/usr/bin/time)	elapsed, user, system	10000
<i>clock()</i>	SO (time.h)	CPU	10000
<i>gettimeofday()</i>	SO (sys/time.h)	elapsed	1
<i>clock_gettime () / clock_getres()</i>	SO (time.h)	elapsed	0.001
<i>omp_get_wtime() / omp_get_wtick()</i>	OpenMP (omp.h)	elapsed	0.001
<i>SYSTEM_CLOCK()</i>	Fortran	elapsed	1

RISC: menos instrucciones que CISC
CISC tienen instrucciones de acceso a memoria de la ALU
Los RISC para acceder a memoria, solamente se pueden usar instrucciones dedicadas a memoria

```
...
for (i=0; i<N; i++) {
    v3[i]=v1[i]+v2[i];
}
...
```

Suma vectores de doubles

```
.L7:
; rax=0, rbx=8N
movsd v1(%rax), %xmm0
addsd v2(%rax), %xmm0
movsd %xmm0, v3(%rax)
addq $8, %rax
cmpq %rbx, %rax
jne .L7
```

i	NI _i	CPI _i
movsd m,r	2N	4
movsd r,m	N	5
addsd m,r	N	1
addq i,r	N	1
cmp r,r	N	1
jne	N	1
	6N	

$T_{CPU} = NI \times CPI \times T_{ciclo}$

$T_{CPU} = NI \times \frac{1}{IPC} \times T_{ciclo}$

$T_{CPU} = \frac{N^{\circ} \text{ciclos}}{\text{código}} \times T_{ciclo}$

$T_{CPU} = \left[\sum_i NI_i \times CPI_i \right] \times T_{ciclo}$

$T_{CPU} = NI \times \left(\frac{\sum_i NI_i \times CPI_i}{NI} \right) \times T_{ciclo}$

$T_{CPU} = NI \times CPI \times T_{ciclo}$

$T_{ciclo} = 1/F$

Para $N = 10^3$ y $F = 100\text{MHz}$ ($\Rightarrow T_{ciclo} = 10^{-8}\text{seg./ciclo}$):

$T_{CPU} \approx \left[6N \times \left(\frac{2N \times 4 + N \times 5 + 3N \times 1}{6N} \right) \right] \times T_{ciclo}$

$= 10^3 \times 16 \text{ ciclos/código} \times 10^{-8}\text{seg./ciclo}$

$= 16 \times 10^{-5}\text{seg./código}$

$$\text{TiempoDeCPU } (T_{CPU}) = \text{CiclosDelPrograma} \cdot T_{CICLO} = \frac{\text{CiclosDelPrograma}}{\text{FrecuenciaDeReloj}}$$

$$\text{CiclosporInstruccion } (CPI) = \frac{\text{CiclosDelPrograma}}{\text{NumeroDeInstrucciones}(NI)}$$

$$T_{CPU} = NI \cdot CPI \cdot T_{CICLO}$$

$$\text{CiclosDelPrograma} = \sum_{i=1}^n CPI_i \cdot I_i$$

$$CPI = \frac{\sum_{i=1}^n CPI_i \cdot I_i}{NI}$$

En el programa hay I_i instrucciones del tipo i ($i=1, \dots, n$).

Cada instrucción del tipo i consume CPI_i ciclos.

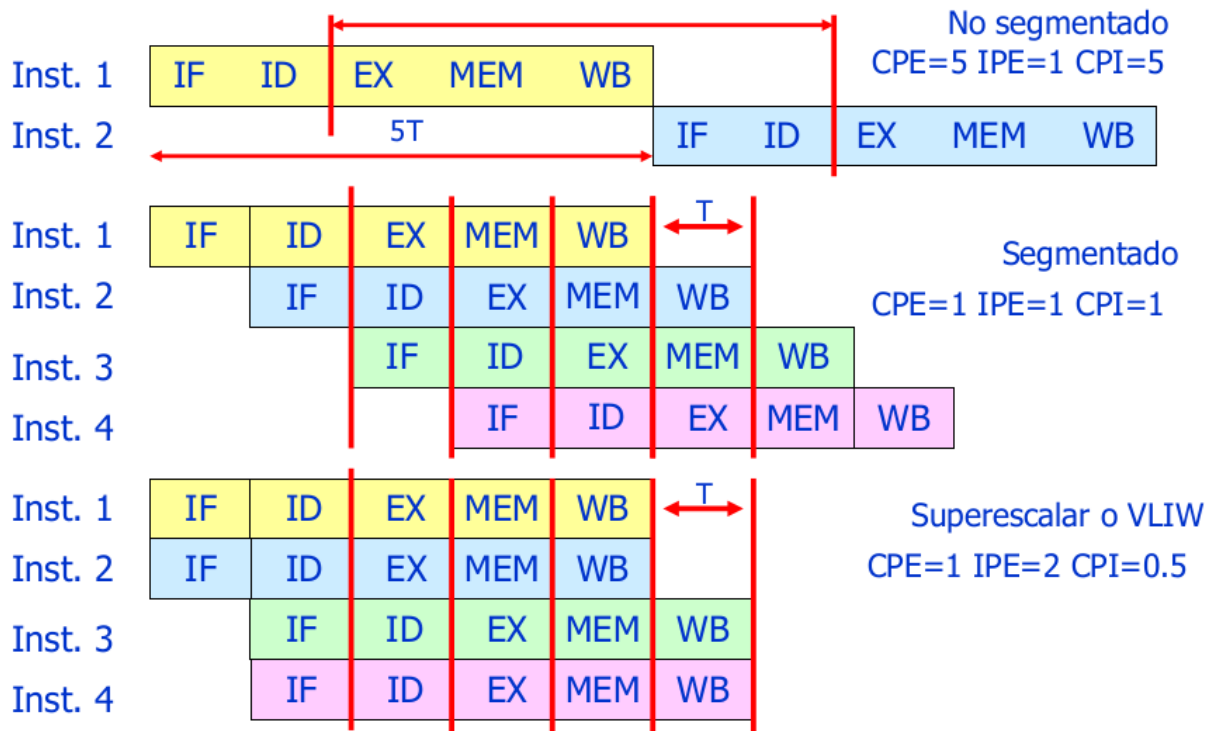
Hay n tipos de instrucciones distintos.

$$T_{CPU} = NI \cdot (CPE/IPE) \cdot T_{CICLO}$$

$$CPI = \frac{CPE}{IPE}$$

Hay procesadores que pueden lanzar para que empiecen a ejecutarse (emitir) varias instrucciones al mismo tiempo.

- **CPE:** Número mínimo de ciclos transcurridos entre los instantes en que el procesador puede emitir instrucciones
- **IPE:** Instrucciones que pueden emitirse (para empezar su ejecución) cada vez que se produce dicha emisión.

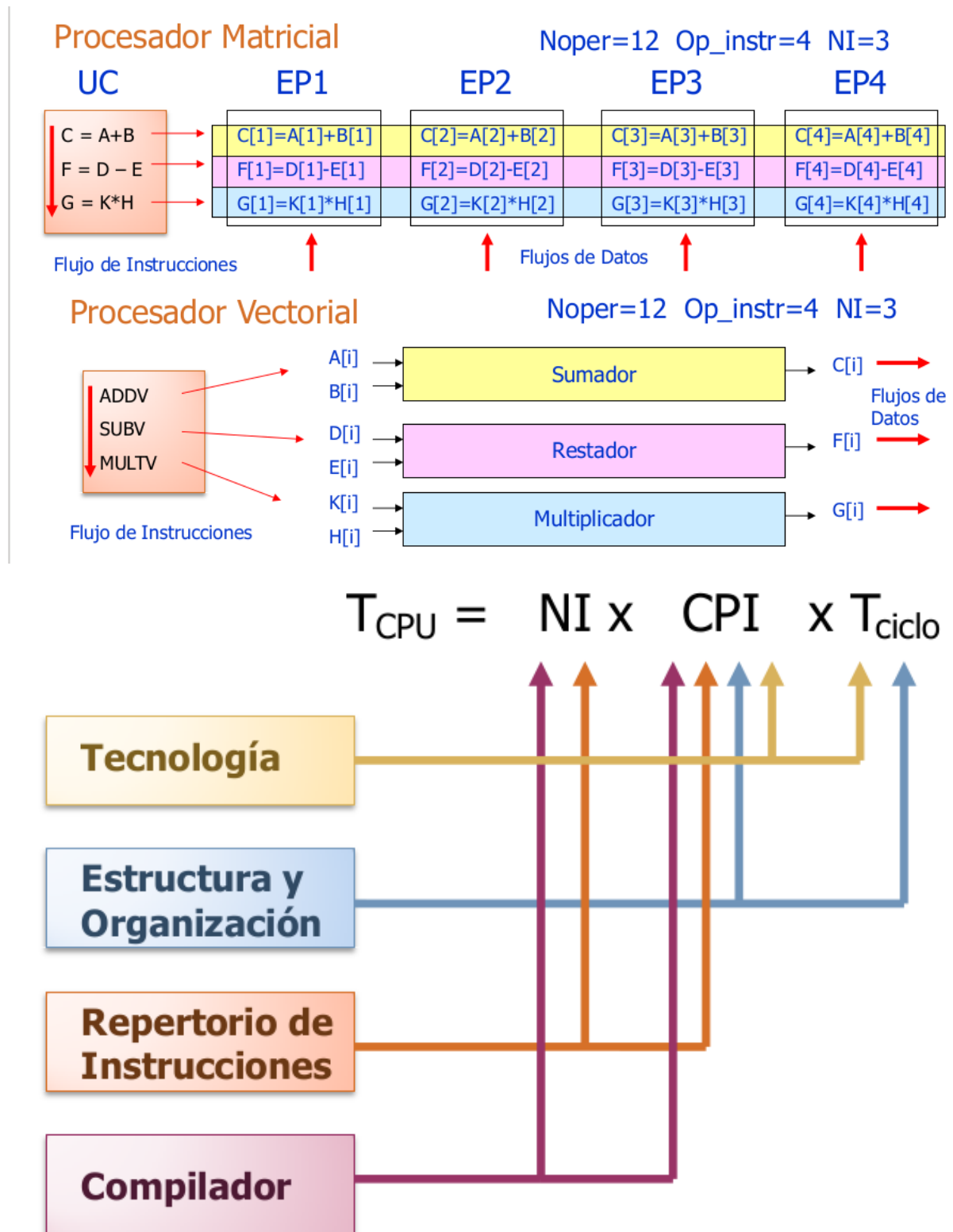


$$T_{CPU} = (N_{oper} / OpInstr) \cdot CPI \cdot T_{CICLO}$$

$$NI = N_{oper} / OpInstr$$

Hay procesadores que pueden codificar varias operaciones en una instrucción.

- **Noper:** Número de operaciones que realiza el programa
- **Op_instr:** Número de operaciones que puede codificar una instrucción.



3.2.2 Productividad: MIPS, MFLOPS.

MIPS: millones de instrucciones por segundo.

$$MIPS = \frac{NI}{T_{CPU} \cdot 10^6} = \frac{F(frecuencia)}{CPI \cdot 10^6}$$

- Depende del repertorio de instrucciones (difícil la comparación de máquinas con repertorios distintos)
- Puede variar con el programa (no sirve para caracterizar la máquina)
- Puede variar inversamente con las prestaciones (mayor valor de MIPS corresponde a peores prestaciones)

MFLOPS: millones de operaciones en coma flotante por segundo.

$$MFLOPS = \frac{\text{OperacionesEnComaFlotante}}{T_{CPU} \cdot 10^6}$$

- No es una medida adecuada para todos los programas (sólo tiene en cuenta las operaciones en coma flotante del programa)
- El conjunto de operaciones en coma flotante no es constante en máquinas diferentes y la potencia de las operaciones en coma flotante no es igual para todas las operaciones (por ejemplo, con diferente precisión, no es igual una suma que una multiplicación..).
- Es necesaria una normalización de las instrucciones en coma flotante

MIPS y FLOPS

```
...
for (i=0; i<N; i++) {
    y[i]=a*x[i]+y[i];
}
...
```

AC ATC

-02

```
;r12=&x,r13=&y, rax=0, rbp=N, xmm1=a
.L6:
    movsd (%r12,%rax,8), %xmm0
    mulsd %xmm1, %xmm0
    addsd (%r13,%rax,8), %xmm0
    movsd %xmm0, (%r13,%rax,8)
    addq $1, %rax
    cmpl %eax, %ebp
    jg .L6
```

T(N=2²⁶)=0.182 seg.

$$GIPS = \frac{NI}{T_{CPU} \times 10^9} = \frac{N \times 7}{0.182 \times 10^9}$$

$$= \frac{2^{26} \times 7}{0.182 \times 10^9} \approx 2.58 \text{ GIPS}$$

$$GFLOPS = \frac{n^o \text{ FP}}{T_{CPU} \times 10^9} = \frac{N \times 2}{0.182 \times 10^9}$$

$$= \frac{2^{26} \times 2}{0.182 \times 10^9} \approx 0.737 \text{ GFLOPS}$$

-03

```
;r12=&x,r13=&y, rax=0, rbp=N/2, xmm1=a
.L7:
    movapd (%r12), %xmm0
    addq $1, %rax
    addq $16, %r12
    addq $16, %r13
    mulpd %xmm1, %xmm0
    addpd -16(%r13), %xmm0
    movaps %xmm0, -16(%r13)
    cmpl %ebp, %eax
    jnb .L7
```

T(N=2²⁶)=0.178 seg.

$$GIPS = \frac{NI}{T_{CPU} \times 10^9} = \frac{(N/2) \times 9}{0.178 \times 10^9}$$

$$= \frac{2^{25} \times 9}{0.178 \times 10^9} \approx 1.7 \text{ GIPS}$$

$$GFLOPS = \frac{2^{26} \times 2}{0.178 \times 10^9} \approx 0.754 \text{ GFLOPS}$$

1.3.3 3.3 Conjunto de programas de prueba (Benchmark).

- Propiedades exigidas a medidas de prestaciones:

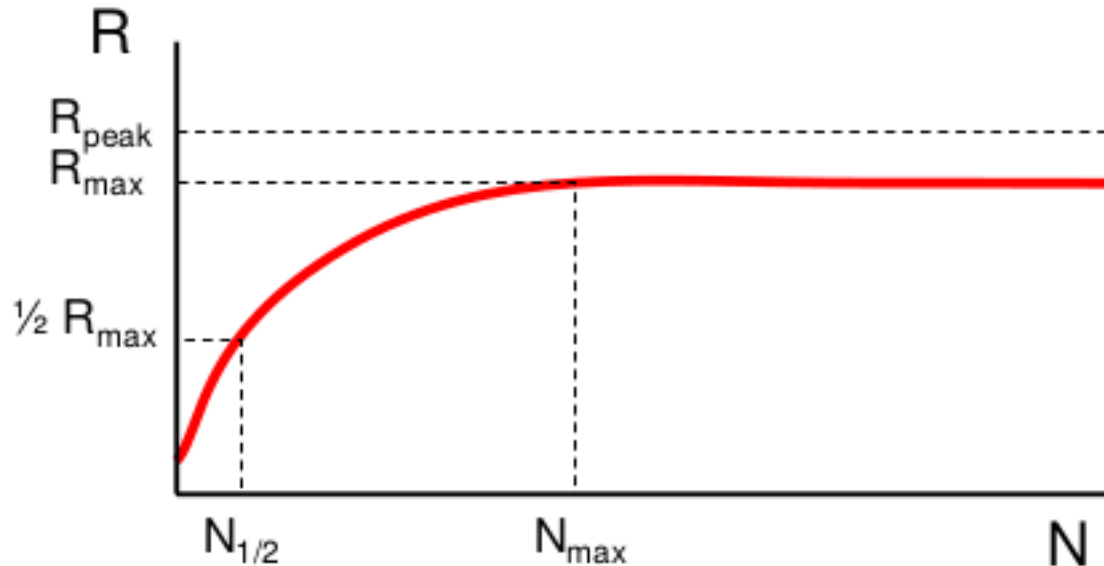
- Fiabilidad => Representativas, evaluar diferentes componentes del sistema y reproducibles
- Permitir comparar diferentes realizaciones de un sistema o diferentes sistemas => Aceptadas por todos los interesados (usuarios, fabricantes, vendedores)
- Interesados:
 - Vendedores y fabricantes de hardware o software.
 - Investigadores de hardware o software.
 - Compradores de hardware o software.
- Tipos de Benchmarks:
 - De bajo nivel o microbenchmark
 - testping-pong, evaluación de las operaciones con enteros o con flotantes.
 - Kernels
 - resolución de sistemas de ecuaciones, multiplicación de matrices, FFT, descomposición LU.
 - Sintéticos
 - Dhrystone, Whetstone.
 - Programas reales.
 - SPEC CPU2006: enteros (gcc, gzip, perlbnk).
 - Aplicaciones diseñadas
 - Predicción de tiempo, simulación de terremotos.

1.3.4 3.3.1 LINPACK.

El núcleo de este programa es una rutina denominada DAXPY (Double precision- real Alpha X Plus Y) que multiplica un vector por una constante y los suma a otro vector. Las prestaciones obtenidas se escalan con el valor de N (tamaño del vector):

```
1 | for (i=0; i<N; i++)
2 |   y[i] = alpha*x[i] + y[i];
```

```
for (i=0 ; i<N ; i++)
  y[i] = alpha*x[i] + y[i];
```



1.3.5 3.4 Ganancia en prestaciones.

3.4.1 Mejora o ganancia de prestaciones (speed-up o ganancia en velocidad).

Si en un computador se incrementan las prestaciones de un recurso haciendo que su velocidad sea p veces mayor (ejemplos: se utilizan p procesadores en lugar de uno, la ALU realiza las operaciones en un tiempo p veces menor. . .):

El incremento de velocidad que se consigue en la nueva situación con respecto a la previa (máquina base) se expresa mediante la ganancia de velocidad o speed-up, S_p

$$S_p = \frac{V_p}{V_1} = \frac{T_1}{T_p}$$

donde

V_1 : velocidad de la máquina base.

V_p : velocidad de la máquina mejorada (un factor p en uno de sus componentes).

T_1 : tiempo de ejecución en la máquina base.

T_p : tiempo de ejecución en la máquina mejorada.

Si se incrementan las prestaciones de un sistema, el incremento en prestaciones (velocidad) que se consigue en la nueva situación, p , con respecto a la previa (sistema base, b) se expresa mediante la ganancia en prestaciones o speed-up, S

$$S = \frac{V_p}{V_b} = \frac{T_b}{T_p}$$

$$S = \frac{T_{CPU}^b}{T_{CPU}^p} = \frac{NI^b \cdot CPI^b \cdot T_{CICLO}^b}{NI^p \cdot CPI^p \cdot T_{ciclo}^p}$$

donde

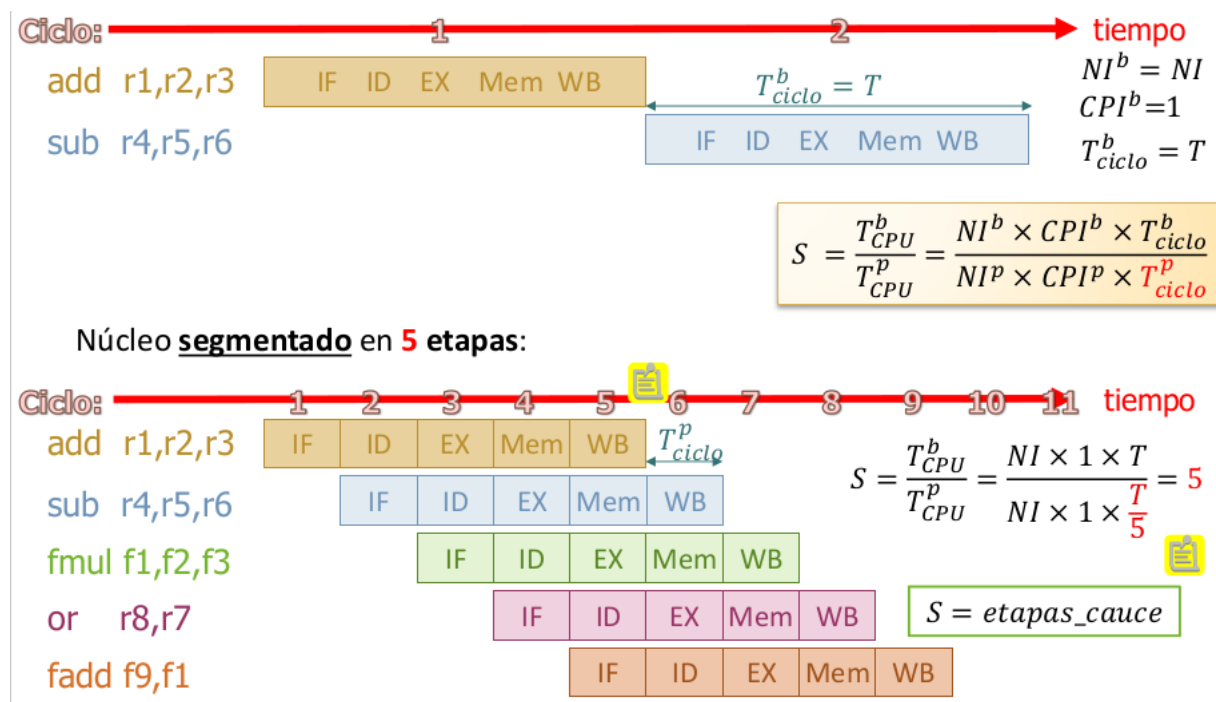
V_b : velocidad de la máquina base.

V_p : velocidad de la máquina mejorada (un factor p en uno de sus componentes).

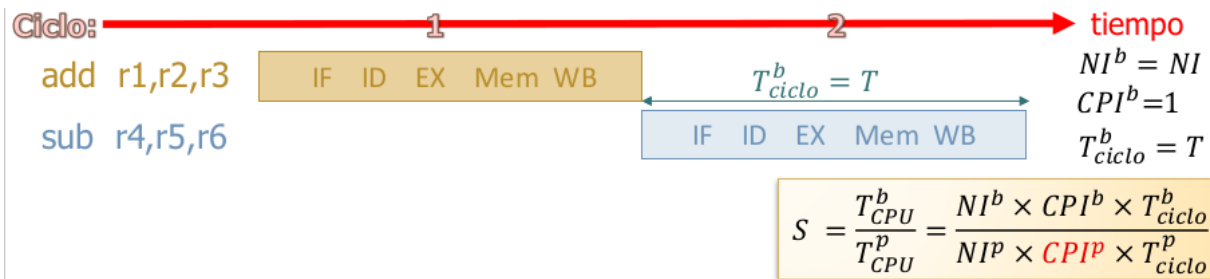
T_b : tiempo de ejecución en la máquina base.

T_p : tiempo de ejecución en la máquina mejorada.

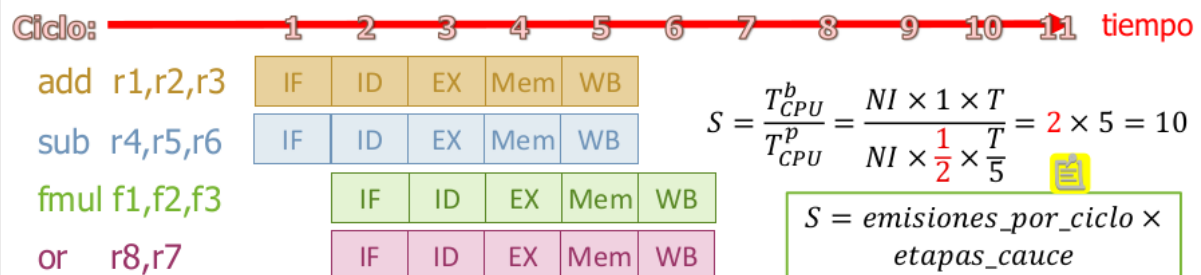
Mejora en un núcleo de procesamiento: segmentación



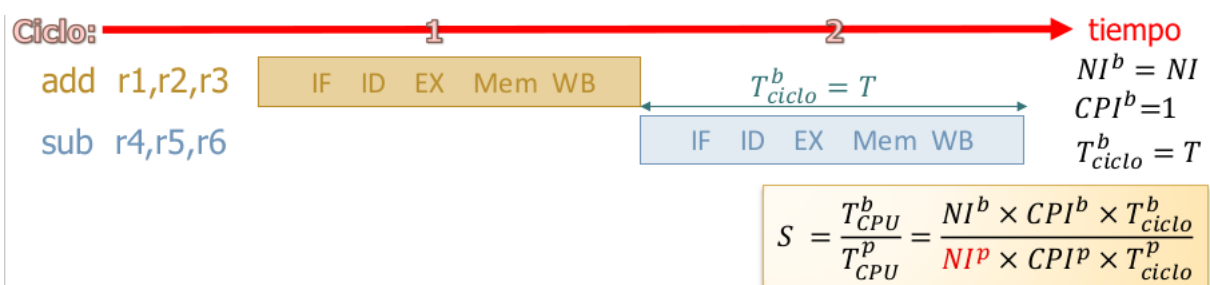
Mejora en un núcleo de procesamiento: operación superescalar



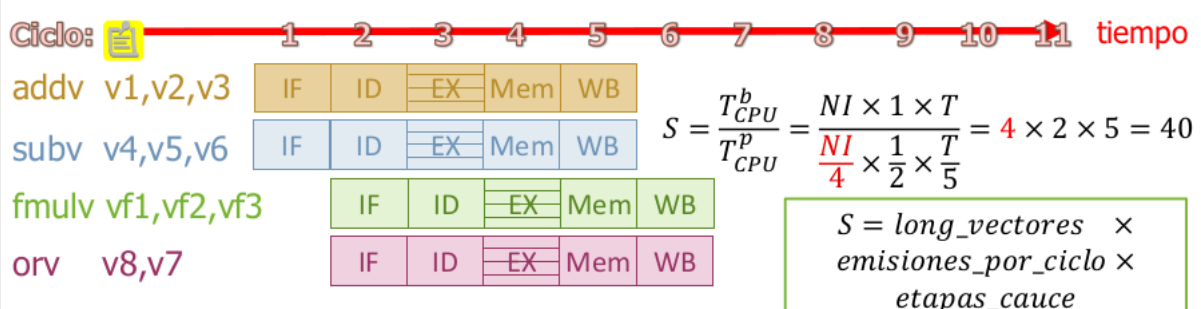
Núcleo **superescalar** con **2 emisiones por ciclo** y **5 etapas**:



Mejora en un núcleo de procesamiento: unidades funcionales SIMD



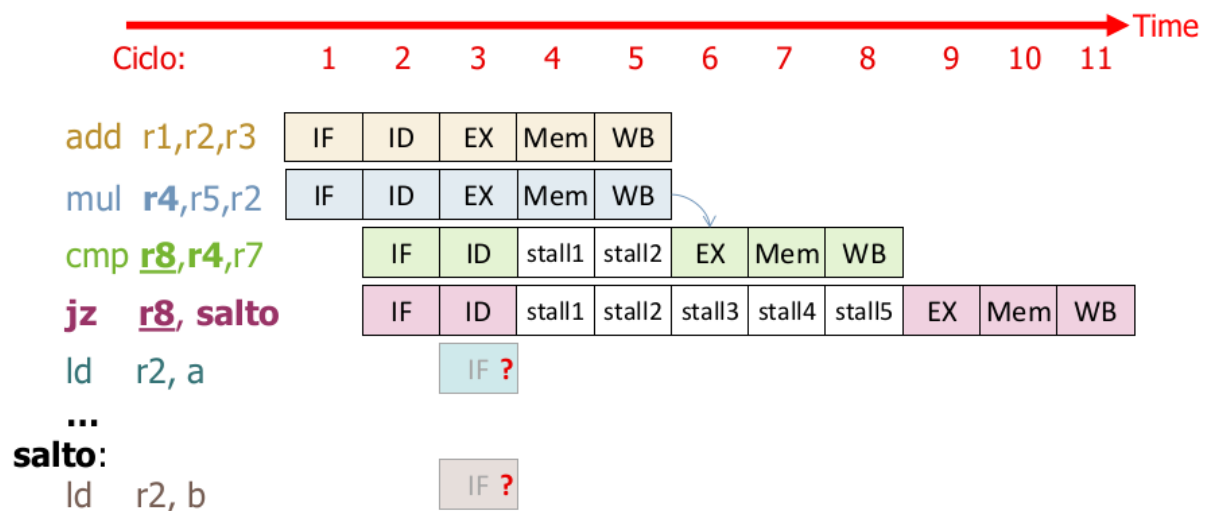
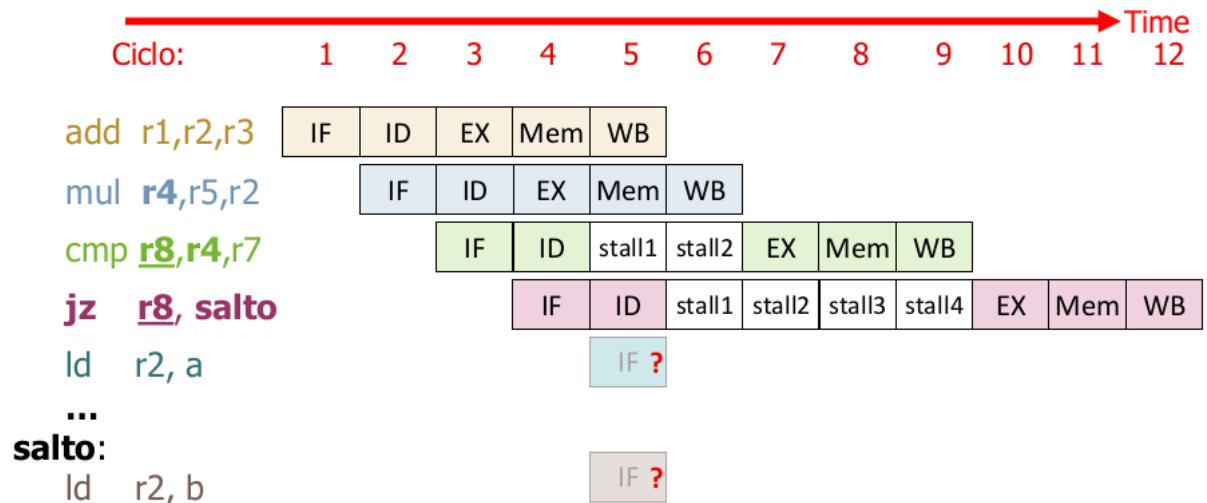
Núcleo **superescalar** con **2 emisiones por ciclo** y **5 etapas**, y **unidades funcionales SIMD** (vectoriales) que procesan **vectores de 4 componentes**:



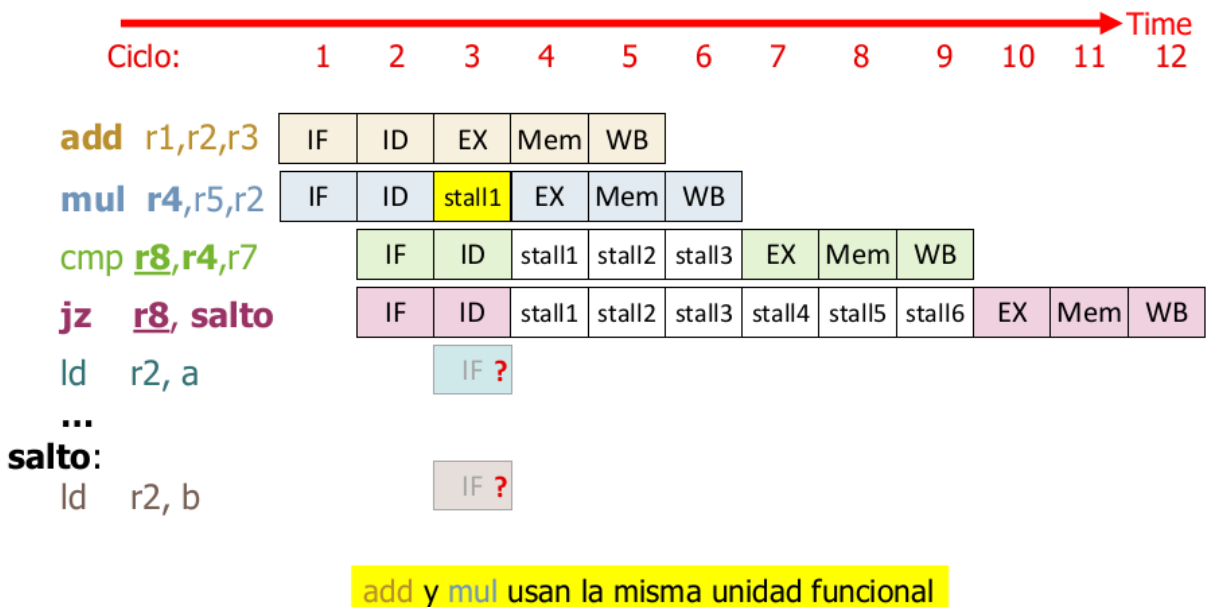
¿Qué impide que se pueda obtener la ganancia en velocidad pico?

- Riesgos:
 - Datos.
 - Control.
 - Estructurales.
- Accesos a memoria (debido a la jerarquía).

Riesgos de datos y control:



Riesgos de datos, control y estructural:



3.4.2 Ley de Amdahl.

La mejora de velocidad, S , que se puede obtener cuando se mejora un recurso de una máquina en un factor p está limitada por:

$$S = \frac{V_p}{V_b} = \frac{T_b}{T_p} \leq \frac{1}{f + \frac{1-f}{p}} = \frac{p}{1+f(p-1)}$$

Si $p \rightarrow \infty$, entonces $\frac{p}{1+f(p-1)} \rightarrow 1/f$.

Si $f \rightarrow 0$, entonces $\frac{p}{1+f(p-1)} \rightarrow p$.

donde f es la fracción del tiempo de ejecución en la máquina sin la mejora durante el que no se puede aplicar esa mejora.

Ejemplo: Si un programa pasa un 25 % de su tiempo de ejecución en una máquina realizando instrucciones de coma flotante, y se mejora la máquina haciendo que estas instrucciones se ejecuten en la mitad de tiempo, entonces $p=2$; $f=0.75$.

$$S \leq 2/(1+0.75)=1.14$$

$$S = \frac{T_b}{T_p} = \frac{1}{0.75 + \frac{0.25}{2}} = 1.14$$

Hay que mejorar el caso más frecuente (lo que más se usa)

Ley enunciada por Amdahl en relación con la eficacia de los computadores paralelos: dado que en un programa hay código secuencial que no puede paralelizarse, los procesadores no se podrían utilizar eficazmente.