Arquitectura de Computadores
Grado en Ingeniería Informática
EII – ULPGC

BLOQUE 4. ARQUITECTURA y PROGRAMACIÓN DE LOS COMPUTADORES PARALELOS

Sumario

- Tema 4-1 Fundamentos del Procesamiento Paralelo
- Tema 4-2 Procesadores de Multihilos Simultáneos
- Tema 4-3 Arquitecturas de los Multiprocesadores de Memoria Compartida
- Tema 4-4 Programación Paralela con OpenMP
- Tema 4-5 Arquitecturas de los Multiprocesadores para Procesamiento Gráfico
- Tema 4-6 Programación Paralela con CUDA
- Tema 4-7 Arquitecturas de los Computadores Paralelos de Paso de Mensajes
- Tema 4-8 Programación Paralela con MPI
- Tema 4-9 Arquitecturas Paralelas Especializadas

Arquitectura de Computadores

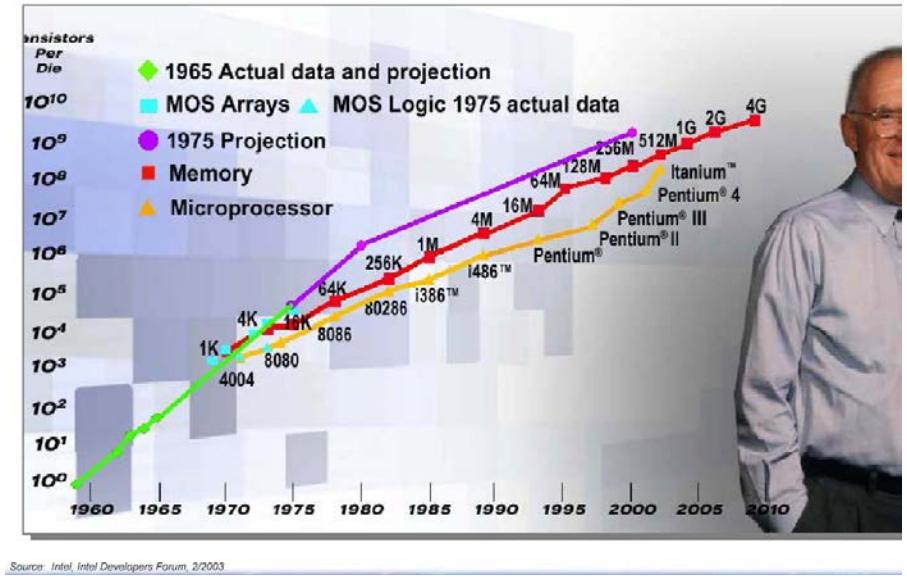
Tema 4-1. Fundamentos del Procesamiento Paralelo

Sumario

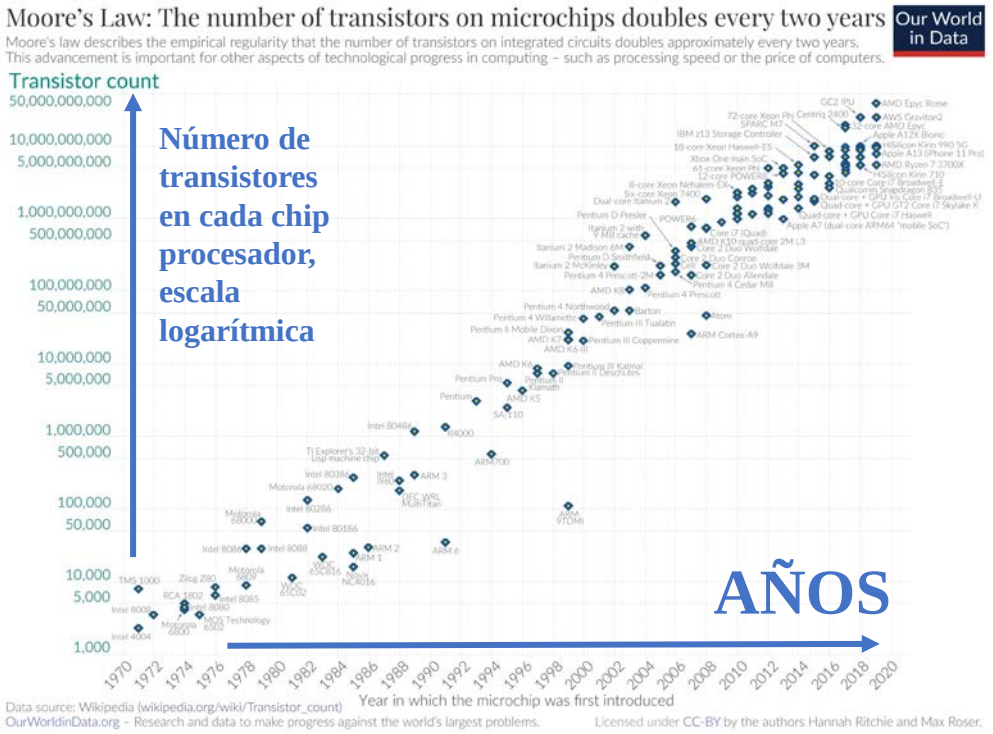
- Introducción
- Terminología
- Ejercicios
- Taxonomía de las arquitecturas paralelas

Introducción

- Limitaciones del aumento de prestaciones de los procesadores
  - La frecuencia de reloj no aumenta significativamente desde 2005
  - La explotación del paralelismo de instrucciones (ILP) no aumenta significativamente
- La ley de Moore sigue cumpliéndose
  - El número de transistores por chip se duplica cada 2 años aproximadamente ☺
  - Este ritmo influye sobre el nivel de prestaciones y el precio de los procesadores
- Los futuros incrementos en prestaciones de los computadores es probable que vayan marcados por la explotación del paralelismo del hardware en vez de por un incremento de la frecuencia de reloj o una reducción del CPI de los procesadores.
- Los programadores que trabajen en las prestaciones de los computadores deben ser programadores paralelos



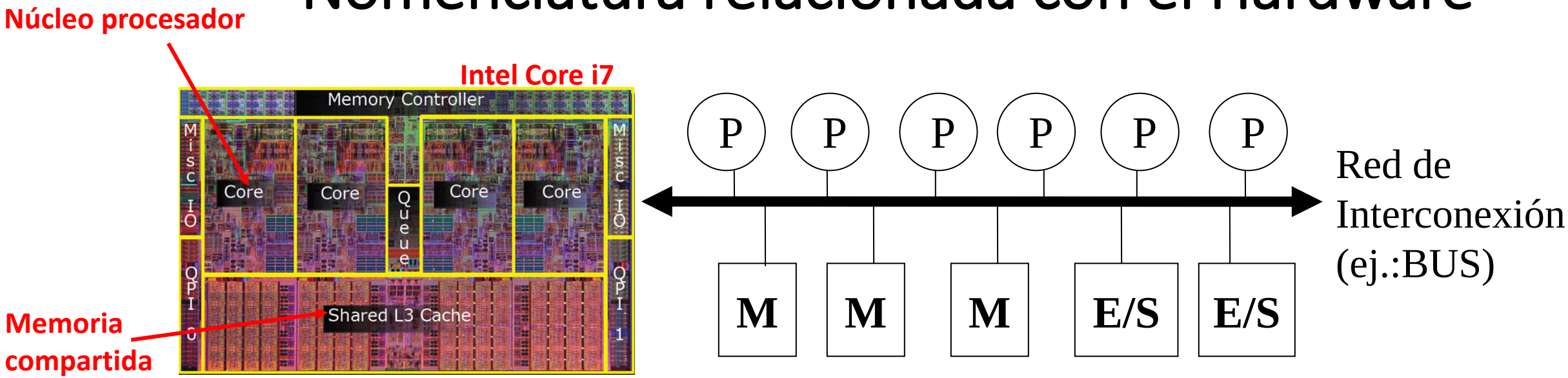
Source: Intel, Intel Developers Forum, 2/2003



Introducción: revolución del Paralelismo

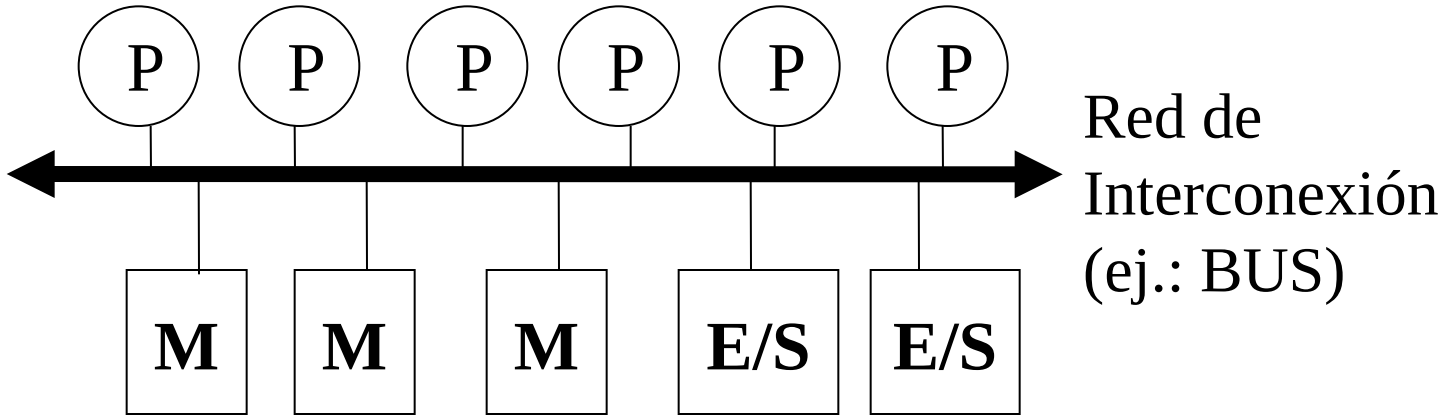
- Computador Paralelo: Camino lógico para aumentar prestaciones de los procesadores
  - Definición: computador con múltiples procesadores conectados entre sí compartiendo recursos de forma coordinada y eficiente.
- El consumo y la disipación de potencia pueden ser un hándicap en el uso de los computadores paralelos, tanto a nivel de multiprocesador como del centro de proceso de datos (CPD, datacenter).
- El software …
  - No es actualmente un obstáculo para explotar los computadores paralelos.
  - Tiene que ser diseñado para usar/adaptarse de forma eficiente a un número variable de procesadores (a esta característica del software se le llama Escalabilidad del Software).
- Innovación proporcionada por la Ingeniería Informática: desarrollar nuevo SW y adaptar el SW existente a los nuevos computadores paralelos comerciales
- Palabras clave: multiprocesador, paralelismo, programa paralelo, clúster, centro de proceso de datos

Nomenclatura relacionada con el Hardware



- *Multiprocesador (multi-núcleo, Inglés: multicore)*: conjunto de al menos dos procesadores que se encuentran integrados en un mismo chip o placa base.
- *Multiprocesador de memoria compartida*: multiprocesador donde los procesadores tienen el mismo espacio de direccionamiento.
  - Extensión natural de los mono-procesadores (un solo núcleo).
- Las prestaciones y/o el ritmo de realización de tareas se aumenta añadiendo procesadores.
- La capacidad de memoria se incrementa, añadiendo módulos de memoria.
- El sistema de E/S se aumenta, añadiendo controladores de E/S.

Nomenclatura relacionada con el Hardware

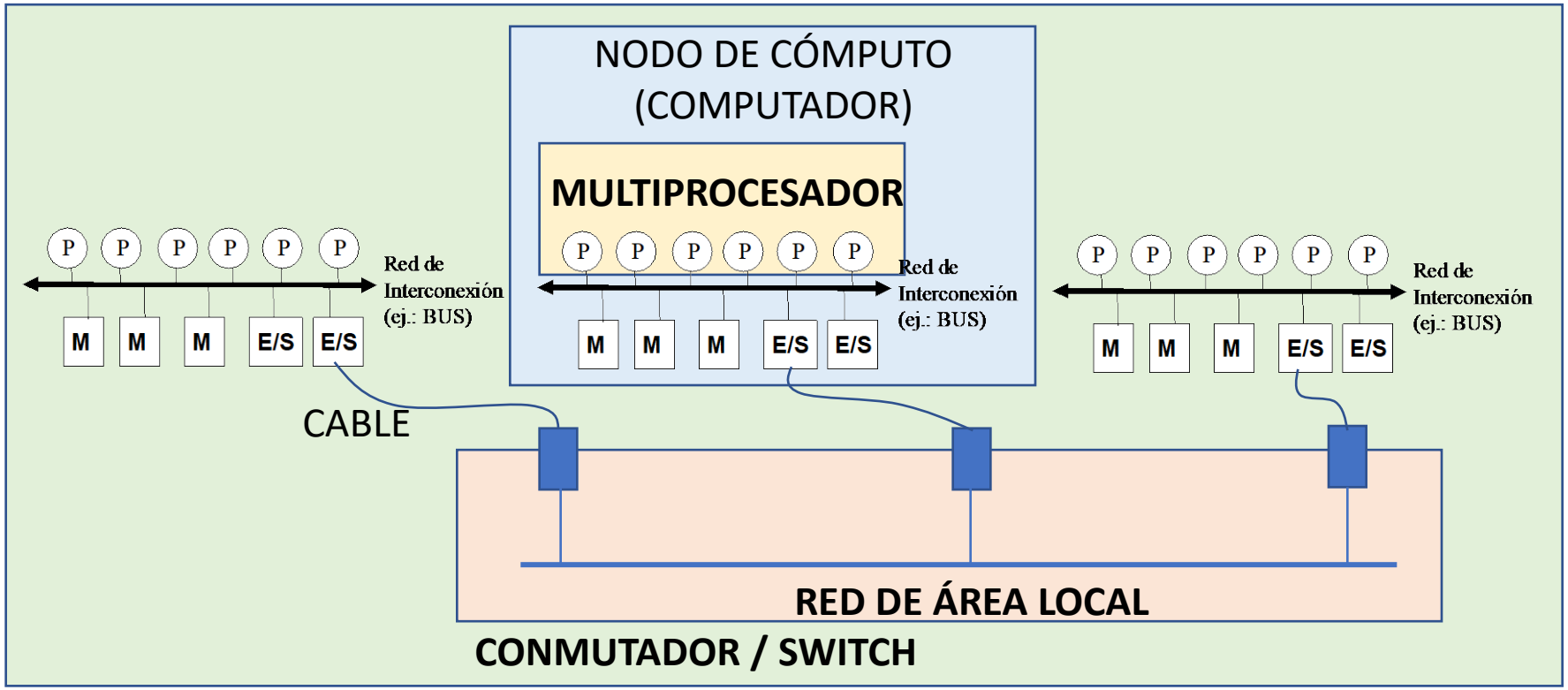


- Tipos de multiprocesadores depende de:
  - Número, tipo y tamaño de los **procesadores.**
  - Número de módulos, tamaño y organización de la **memoria.**
  - Tipo de **interconexión** entre los procesadores y la memoria.
    - Estática (no modificable, ejemplo: bus), Dinámica (ejemplo: red interconexión en chip).
    - Procesadores acoplados débilmente (red de interconexión), acoplados fuertemente (bus).

Nomenclatura relacionada con el Hardware

19”: 48,26 cm

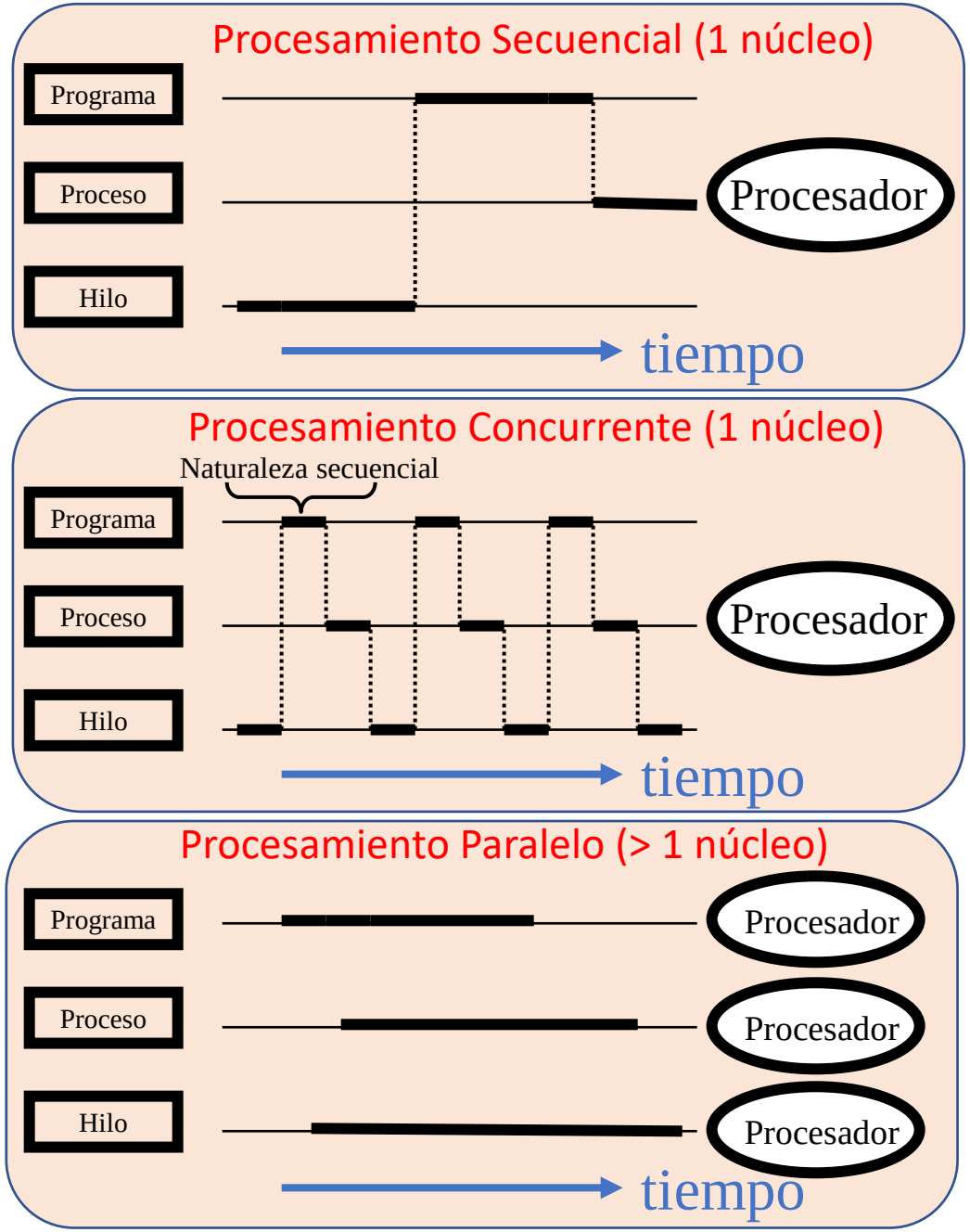
CLÚSTER



- Clúster (multicomputador): conjunto de computadores conectados a través de una red de área local que funciona como si hubiera un multiprocesador grande. **Cada computador utiliza un espacio de direccionamiento distinto**.

Nomenclatura relacionada con el Software

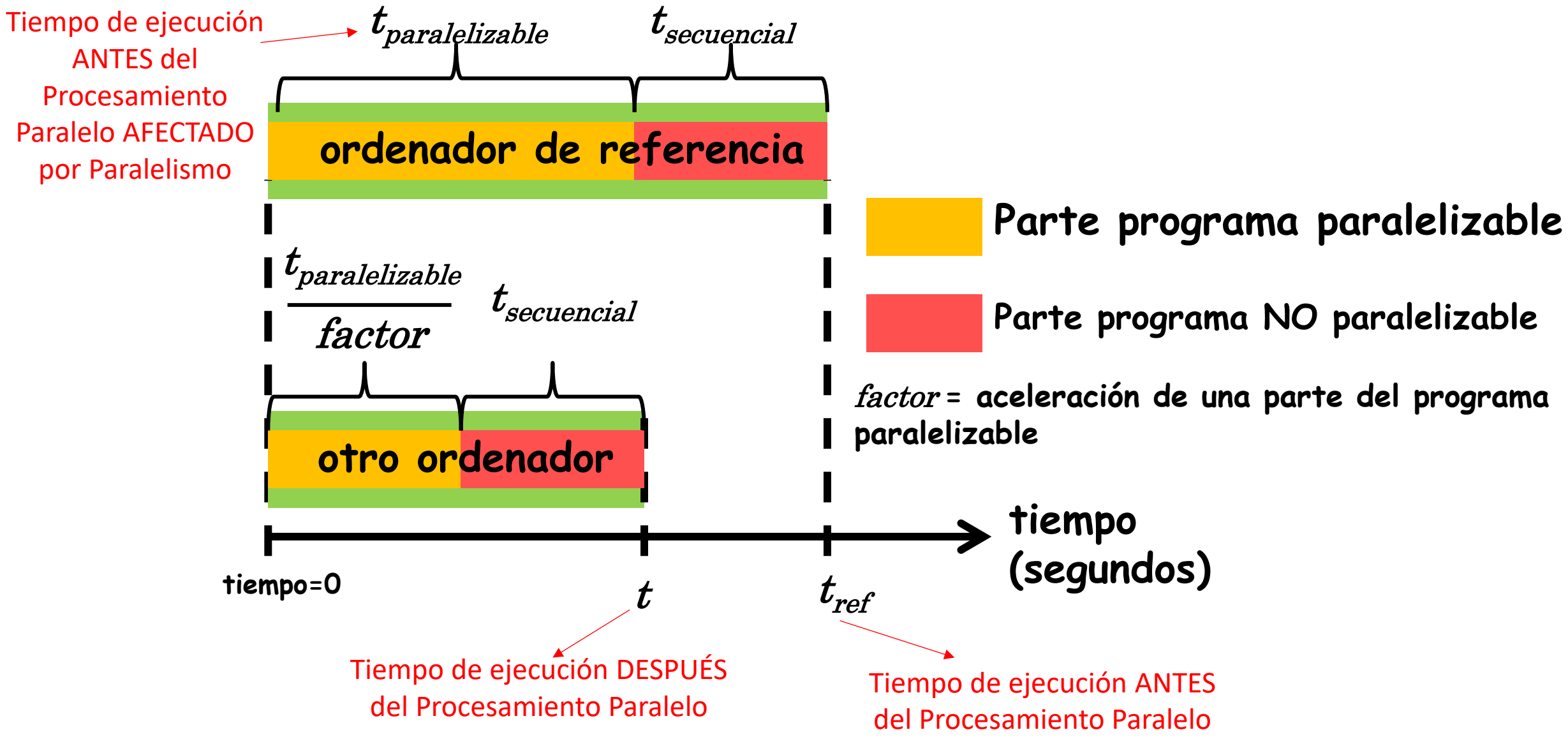
- Programa: fichero ejecutable que contiene un conjunto de códigos de instrucción que se combinan para realizar un determinado trabajo en un ordenador.
- Proceso: la activación del programa en un momento dado; puede estar dividido en hilos; los procesos no comparten el mismo subespacio de direccionamiento.
- Hilo (thread): conjunto de instrucciones independientes que comparten un mismo espacio de direccionamiento.
- Procesamiento Secuencial: programas hacen uso de un solo procesador desde el inicio al final del programa.
- Procesamiento Concurrente: programas hacen uso de uno solo procesador durante un intervalo de tiempo que es inferior al tiempo que tarda todo el programa.
- Procesamiento Paralelo: programas hacen uso en cada instante de tiempo de varios de los procesadores disponibles en el hardware.



Tipos y niveles del Procesamiento Paralelo

- Tipos de Procesamiento Paralelo o Paralelismo
  - Funcional: se aplica a la solución algorítmica de un problema y se caracteriza por ser irregular a excepción de los bucles anidados.
  - Datos: se aplica a las estructuras de datos (vectores, matrices, etc.) y se caracteriza por ser muy regular como en las aplicaciones numéricas y multimedia.
- Niveles de implementación software del Paralelismo
  - Instrucciones: diferentes instrucciones de un programa se ejecutan en distintas UFs (instruction-level parallelism).
  - Hilos: varios hilos independientes de instrucciones que comparten el mismo espacio de direccionamiento se ejecutan en distintos procesadores físicos/lógicos.
  - Procesos: varios programas/procedimientos/tareas independientes se ejecutan simultáneamente en paralelo ({task, process}-level parallelism) en distintos procesadores.
  - Programa: varios programas se ejecutan simultáneamente en paralelo en varios procesadores (parallel processing program).
- Técnicas hardware básicas
  - Segmentación
  - Replicación

Ejercicios


Tiempo de ejecución ANTES del Procesamiento Paralelo AFECTADO por Paralelismo
$t_{paralelizable}$
$t_{secuencial}$
ordenador de referencia
Parte programa paralelizable
Parte programa NO paralelizable
$\frac{t_{paralelizable}}{factor}$
$t_{secuencial}$
factor = aceleración de una parte del programa paralelizable
otro ordenador
tiempo (segundos)
tiempo=0
t
t_{ref}
Tiempo de ejecución DESPUÉS del Procesamiento Paralelo
Tiempo de ejecución ANTES del Procesamiento Paralelo

Ejercicio 1: ley de Amdahl

- Suponiendo que se quiere conseguir un speed-up de 90X usado 100 procesadores. ¿Qué porcentaje del tiempo de la computación original puede ser secuencial?
- Solución:

$$\text{Tiempo de ejecución DESPUÉS del procesamiento paralelo} = \frac{\text{Tiempo de ejecución ANTES del Procesamiento Paralelo AFECTADO por Paralelismo}}{\text{Factor de mejora}} + \text{Tiempo de ejecución ANTES del Procesamiento Paralelo NO AFECTADO por Paralelismo}$$

$$\text{Tiempo de ejecución DESPUÉS del procesamiento paralelo} = \frac{\text{Tiempo de ejecución ANTES del Procesamiento Paralelo AFECTADO por Paralelismo}}{\text{Factor de mejora}} + \text{Tiempo de ejecución ANTES del Procesamiento Paralelo} - \text{Tiempo de ejecución ANTES del Procesamiento Paralelo AFECTADO por Paralelismo}$$

Ejercicio 1 - solución

$$\text{Speed-Up} = \frac{\text{Tiempo de ejecución ANTES del Procesamiento Paralelo}}{\text{Tiempo de ejecución DESPUÉS del Procesamiento Paralelo}}$$

$$= \frac{\text{Tiempo de ejecución ANTES del procesamiento paralelo}}{\frac{\text{Tiempo de ejecución ANTES del Procesamiento Paralelo AFECTADO por Paralelismo}}{\text{Factor de mejora}} + \text{Tiempo de ejecución ANTES del Procesamiento Paralelo} - \text{Tiempo de ejecución ANTES del Procesamiento Paralelo AFECTADO por Paralelismo}}$$

$$= \frac{1}{\frac{\text{Fracción de tiempo de ejecución ANTES de P.P. AFECTADO por Paralelismo}}{\text{Factor de mejora}} + 1 - \text{Fracción de tiempo de ejecución ANTES de P.P. AFECTADO por Paralelismo}}$$

$$90 = \frac{1}{\frac{\text{Fracción de tiempo de ejecución ANTES de P.P. AFECTADO por Paralelismo}}{100} + 1 - \text{Fracción de tiempo de ejecución ANTES de P.P. AFECTADO por Paralelismo}}$$

Fracción de tiempo de ejecución ANTES de P.P. AFECTADO por Paralelismo = 0,9989

Fracción de tiempo de ejecución ANTES de P.P. NO afectado por Paralelismo (procesamiento secuencial) = $1 - 0{,}9989 = 1{,}122\ 10^{-3}$ (0,1122 %)

Ejercicio 2: influencia de la carga computacional

- Suponer que se quieren realizar dos sumas. La primera es la suma de 10 números. La segunda consiste en la suma de dos matrices 2-D cuyas dimensiones son 10 x 10. Suponer que sólo la suma matricial es paralelizable. ¿Cuál es el speed-up cuando se pasa de 10 a 40 procesadores? ¿Cuál es el speed-up cuando las matrices son 20 x 20?
- Solución: 10 números (secuencial) + 2 matrices de 10 x 10 (paralelo)

$$\text{Tiempo de ejecución DESPUÉS del Procesamiento Paralelo} = \frac{\text{Tiempo de ejecución ANTES de P.P. AFECTADO por Paralelismo}}{\text{Factor de mejora}} + \text{Tiempo de ejecución ANTES de P.P. NO AFECTADO por Paralelismo}$$

$$\text{Tiempo de ejecución DESPUÉS del Procesamiento Paralelo (10 P)} = \frac{100\ t}{10} + 10\ t = 20\ t$$

(t: tiempo de una suma)

$$\text{Speed-Up (10 P)} = \frac{\text{Tiempo de ejecución ANTES del Procesamiento Paralelo}}{\text{Tiempo de ejecución DESPUÉS del Procesamiento Paralelo}} = \frac{(100 + 10)\ t}{20\ t} = 5{,}5X$$

$$\textbf{Eficiencia Paralelismo (10 P)} = 100\% \times \frac{\text{Speed-Up (10 P)}}{\text{Número de procesadores (10)}} = \frac{100\% \times 5{,}5X}{10} = \textbf{55\%}$$

Ejercicio 2

$$\text{Tiempo de ejecución DESPUÉS del Procesamiento Paralelo (40 P)} = \frac{100\ t}{40} + 10\ t = 12{,}5\ t$$

$$\text{Speed-Up (40 P)} = \frac{\text{Tiempo de ejecución ANTES del Procesamiento Paralelo}}{\text{Tiempo de ejecución DESPUÉS del Procesamiento Paralelo}} = \frac{(100 + 10)\ t}{12{,}5\ t} = 8{,}8X$$

$$\textbf{Eficiencia Paralelismo (40 P)} = 100\% \times \frac{\text{Speed-Up (40 P)}}{\text{Número de procesadores (40)}} = \frac{100\% \times 8{,}8X}{40} = \textbf{22,0\%}$$

- Solución: 10 números + 2 matrices de 20 x 20

$$\text{Tiempo (10 P)} = \frac{400\ t}{10} + 10\ t = 50\ t \qquad \text{Speed-Up (10 P)} = \frac{(400 + 10)\ t}{50\ t} = 8{,}2X \qquad \text{Eficiencia Paralelismo (10 P)} = \frac{100\% \times 8{,}2X}{10} = \textbf{82,0\%}$$

$$\text{Tiempo (40 P)} = \frac{400\ t}{40} + 10\ t = 20\ t \qquad \text{Speed-Up (40 P)} = \frac{(400 + 10)\ t}{20\ t} = 20{,}5X \qquad \text{Eficiencia Paralelismo (40 P)} = \frac{100\% \times 20{,}5X}{40} = \textbf{51,3\%}$$

Concepto de Eficiencia del Paralelismo

- Resultados del Ejercicio 2 (anterior)

Carga computacional (410t): parte paralelizable y parte no-paralelizable

$$\text{Tiempo (10 P)} = \frac{400\ t}{10} + 10\ t = 50\ t$$

$$\text{Speed-Up (10 P)} = \frac{(400 + 10)\ t}{50\ t} = 8{,}2X$$

$$\text{Eficiencia Paralelismo (10 P)} = \frac{100\% \times 8{,}2X}{10} = \mathbf{82{,}0\%}$$

10 procesadores (P) reales

$$\text{Tiempo (}\mathbf{8{,}2\ P\ \underline{equivalentes}}\text{)} = \frac{(400 + 10)\ t}{8{,}2X} = 50\ t$$

82,0% de los procesadores disponibles operan continuamente

CONCEPTO: Eficiencia del Paralelismo representa el **porcentaje de procesadores disponibles** que están continuamente funcionando durante la ejecución del programa suponiendo que toda la carga computacional es paralelizable

Nomenclatura relacionada con el software para Procesamiento Paralelo

- Escalar el software paralelo: **aumentar el número de procesadores** disponibles para ejecutar el programa paralelo o **aumentar el volumen de datos** que son procesados por el programa paralelo.
- Escalado fuerte: aumenta el número de procesadores manteniendo el tamaño del problema fijo (se fija el volumen de datos procesados).
- Escalado débil: aumenta el tamaño del problema (aumenta el volumen de datos procesados) de forma proporcional al aumento del número de procesadores.

Ejercicio 3: influencia del BALANCEO DE CARGA

- En el Ejercicio 2 se supuso que la carga computacional estaba perfectamente balanceada entre los 40 procesadores. De esta forma, se consiguió un speed-up de 20,5X cuando las matrices son 20 x 20. Esto significa que cada procesador se le asignó el 2,5% (100%/40) del trabajo a realizar. En vez de esta situación ideal, suponer que a un procesador se le ha asignado mayor cantidad de carga computacional que a los demás. Calcular el impacto sobre el speed-up si esta carga es del 5% o del 12,5% para matrices de 20 x 20. ¿Cuál es el porcentaje de tiempo en el que el resto de los procesadores están siendo utilizados?
- Solución: matrices de 20 x 20, desbalanceo: 5%, a un procesador se le asigna el doble de trabajo que se le asignaría de forma equitativa

Carga paralelizable: 400 t 1 P: 5% Carga paralelizable: 20 t 39 P: 95% Carga paralelizable: 380 t

Tiempo de ejecución DESPUÉS del Procesamiento Paralelo (40 P) $= \text{Max}\left(\frac{20\text{ t}}{1}, \frac{380\text{ t}}{39}\right) + 10\text{ t} = 30\text{ t}$

Speed-Up (40 P) $= \frac{(400 + 10)\text{ t}}{30\text{ t}} = 13{,}7\text{X}$ **Empeoramiento: 20,5X → 13,7X**

33,2% de reducción de prestaciones en el multiprocesador

Porcentaje de tiempo en el que el resto de los procesadores están siendo utilizados $= 100\% \times \frac{\frac{380\text{ t}}{39}}{\frac{20\text{ t}}{1}} = 48{,}7\%$

La mitad del tiempo, el 97% de procesadores (39/40) están siendo inutilizados!! 😔 😔 😔

Ejercicio 3

- Solución: matrices de 20 x 20, desbalanceo: 12,5%

Carga paralelizable: 400 t 1P: 12,5% Carga paralelizable: 50 t 39 P: 87,5% Carga paralelizable: 350 t

Tiempo de ejecución DESPUÉS del procesamiento paralelo (40 P) $= \text{Max}\left(\frac{50\text{ t}}{1}, \frac{350\text{ t}}{39}\right) + 10\text{ t} = 60\text{ t}$

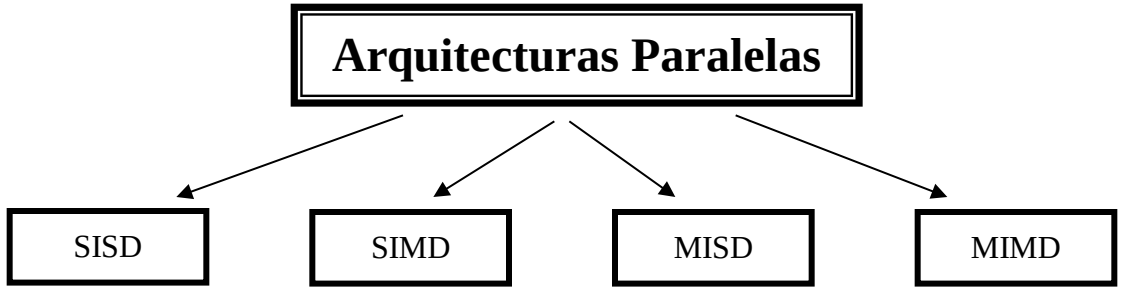
Speed-Up (40 P) $= \frac{(400 + 10)\text{ t}}{60\text{ t}} = 6{,}8\text{X}$ **Empeoramiento: 20,5X → 6,8X**

66,8% de reducción de prestaciones en el multiprocesador

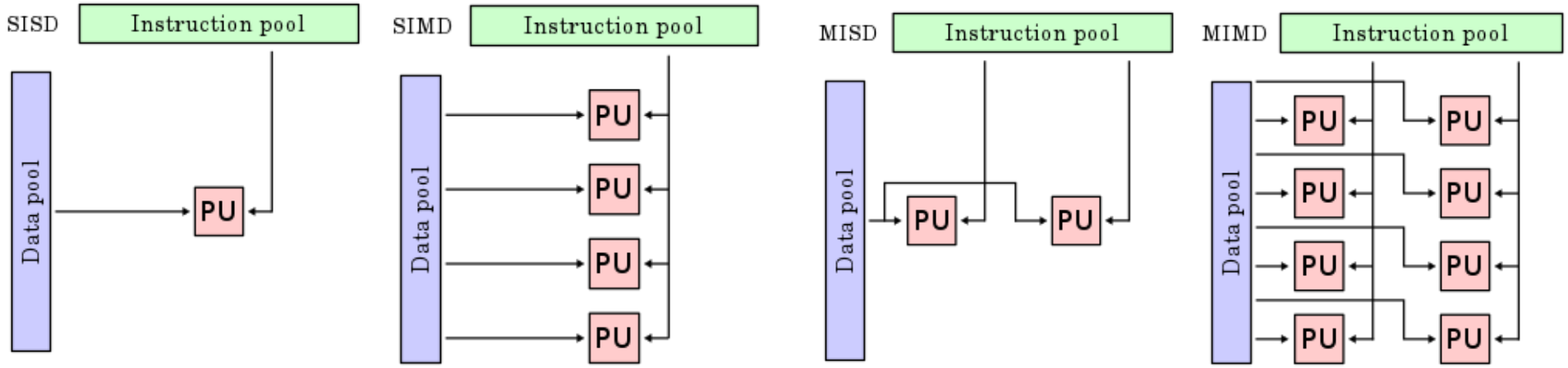
Porcentaje de tiempo en el que el resto de los procesadores están siendo utilizados $= 100\% \times \frac{\frac{350\text{ t}}{39}}{\frac{50\text{ t}}{1}} = 17{,}9\%$

Más de 4/5 partes del tiempo (82,1%), el 97% de procesadores (39/40) están siendo inutilizados!! 😔 😔 😔 😔

Clasificación de Arquitecturas de los Computadores Paralelos (HW)



Taxonomía de Flynn



Clasificación de de Arquitecturas los Computadores Paralelos (HW)