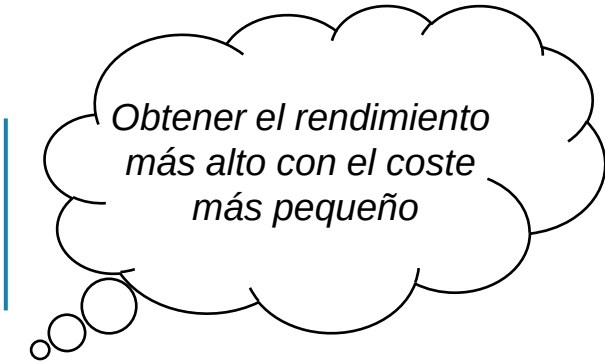
INTRODUCCIÓN A LA EVALUACIÓN DEL RENDIMIENTO

Usuarios, administradores
y diseñadores



RENDIMIENTO

Incluso en los sistemas informáticos más comunes en nuestra vida diaria, como los ordenadores personales, tabletas y teléfonos móviles, el buen rendimiento de los mismos es esencial.

Como usuarios, nos preocupa si nuestro ordenador es lento, lo cual es resultado de una valoración un tanto subjetiva pero basada en algo tan claro y medible como es el tiempo que tarda en ejecutar las tareas, programas o servicios.

- Esas tareas constituyen la demanda de trabajo que solicitamos al sistema informático y se conoce como carga de trabajo (workload).
- Esa demanda de trabajo para los sistemas solicita distintos recursos, como por ejemplo para el caso de nuestro ordenador personal, el disco duro, el procesador o la memoria, entre otros.

CARGA, EVALUACIÓN Y MEDIDA

La carga de trabajo (*workload*) es el conjunto de tareas que debe hacer un sistema, en un momento determinado.

- Esta carga puede ser tan grande o inmanejable que se suele tomar un subconjunto como carga (carga de prueba o de test) para poder evaluar su rendimiento.
- La carga de trabajo posee unas variables que reflejan su comportamiento cuantitativo, por ejemplo, el número de accesos a una página web.

CARGA, EVALUACIÓN Y MEDIDA

La evaluación del rendimiento, a grandes rasgos, supone que una carga de trabajo (o de prueba) es sometida a un sistema, con una determinada configuración.

- Al funcionar el sistema con esa carga, también se pueden obtener los valores de ciertas magnitudes de variables predefinidas, o medidas cuantitativas del sistema
  - Por ejemplo, en un servidor web (sistema) sometido a las transacciones que recibe/envía a los usuarios (carga) se podría establecer su rendimiento en número medio de transacciones/s, midiendo la cantidad de esas transacciones *http* en un periodo de tiempo.

MEDIDAS DE RENDIMIENTO

Tiempo de respuesta (*response time*)

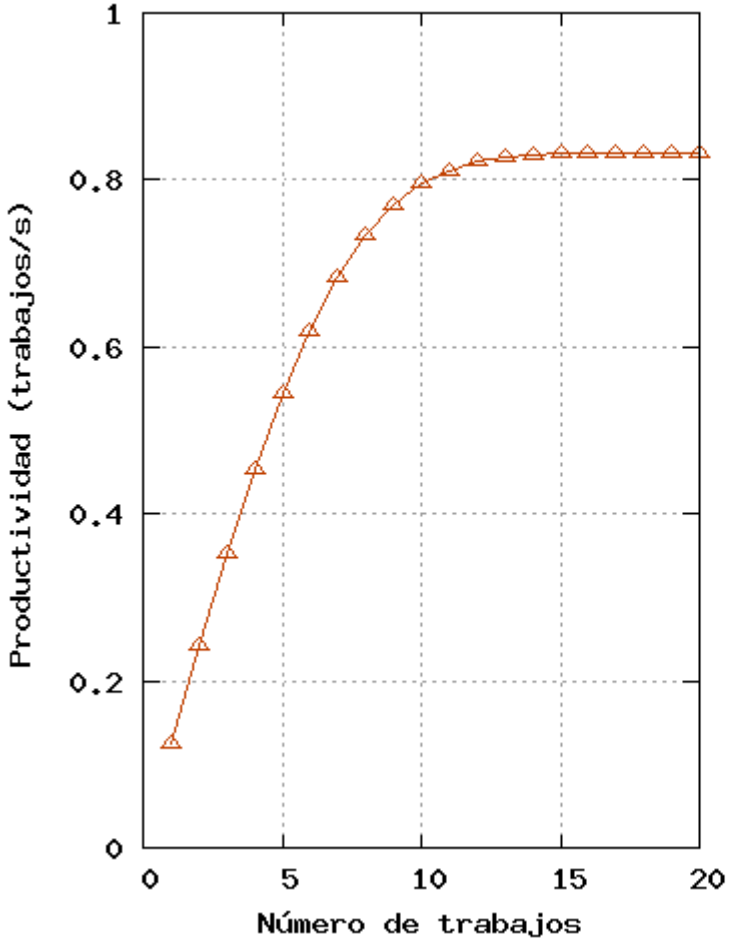
- Tiempo total desde el principio hasta el final de la actividad
  - Tiempo de ejecución de un programa (s)
  - Tiempo de acceso a un disco (ms)

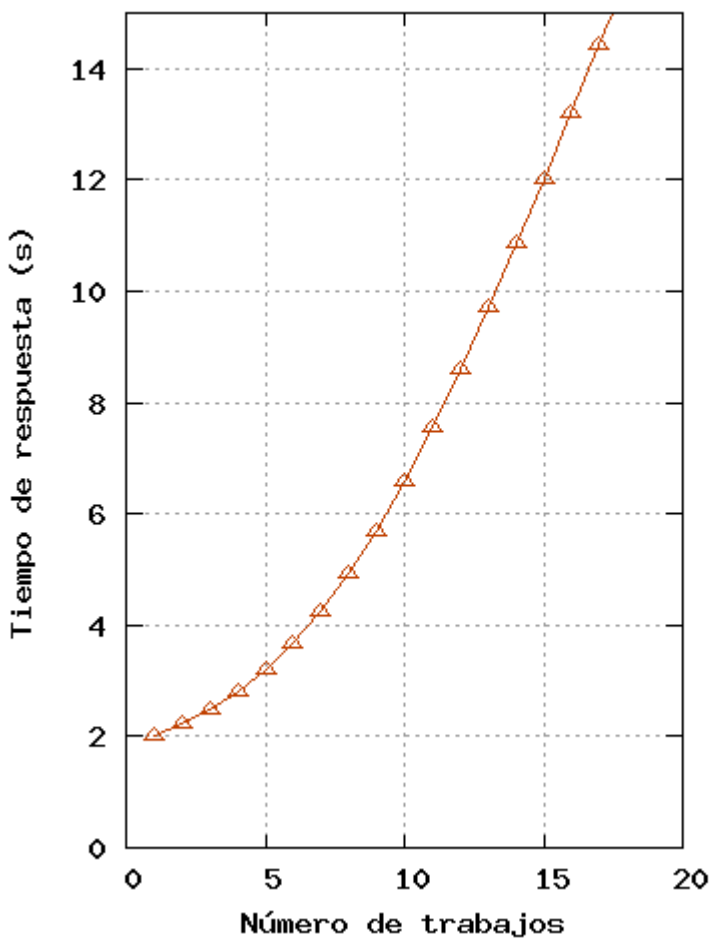
¡La más fiable e intuitiva para comparar rendimientos!

Productividad (*throughput*)

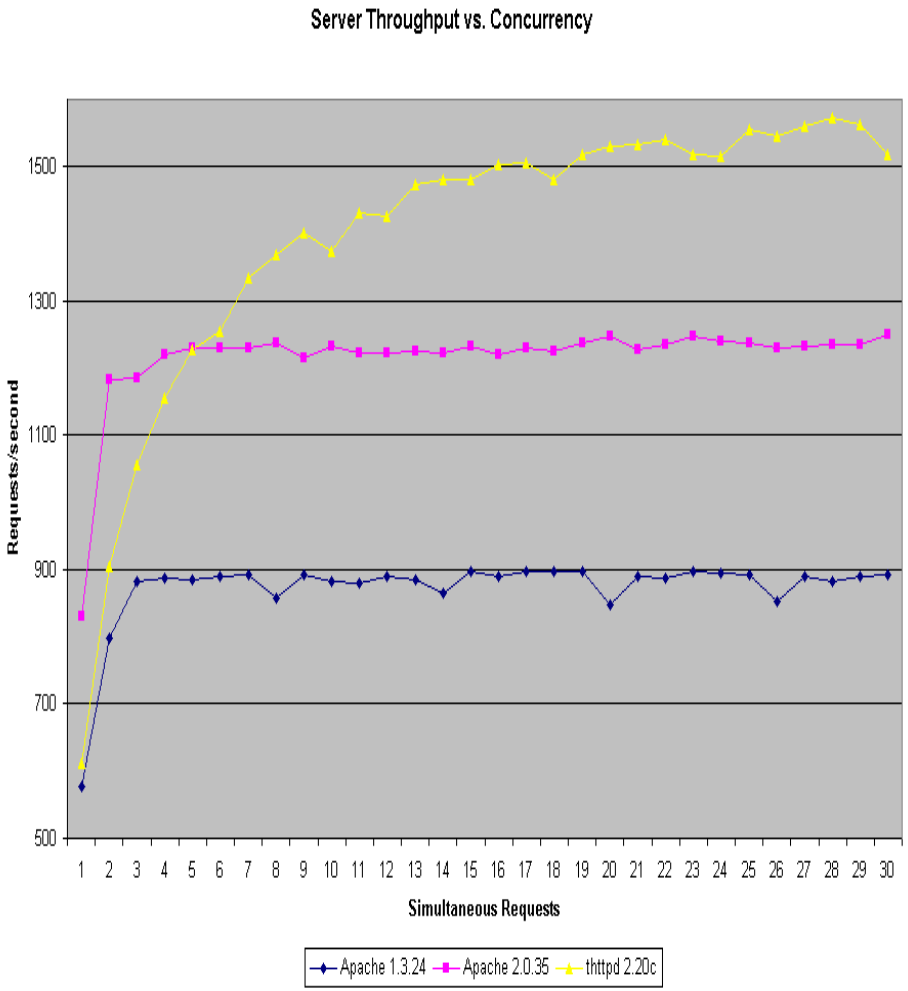
- Cantidad de trabajo hecho por unidad de tiempo
  - Programas ejecutados por hora
  - Páginas por hora servidas por un servidor web
  - Correos por segundo procesados por un servidor de correo
  - Peticiones por minuto procesados por un servidor de comercio electrónico

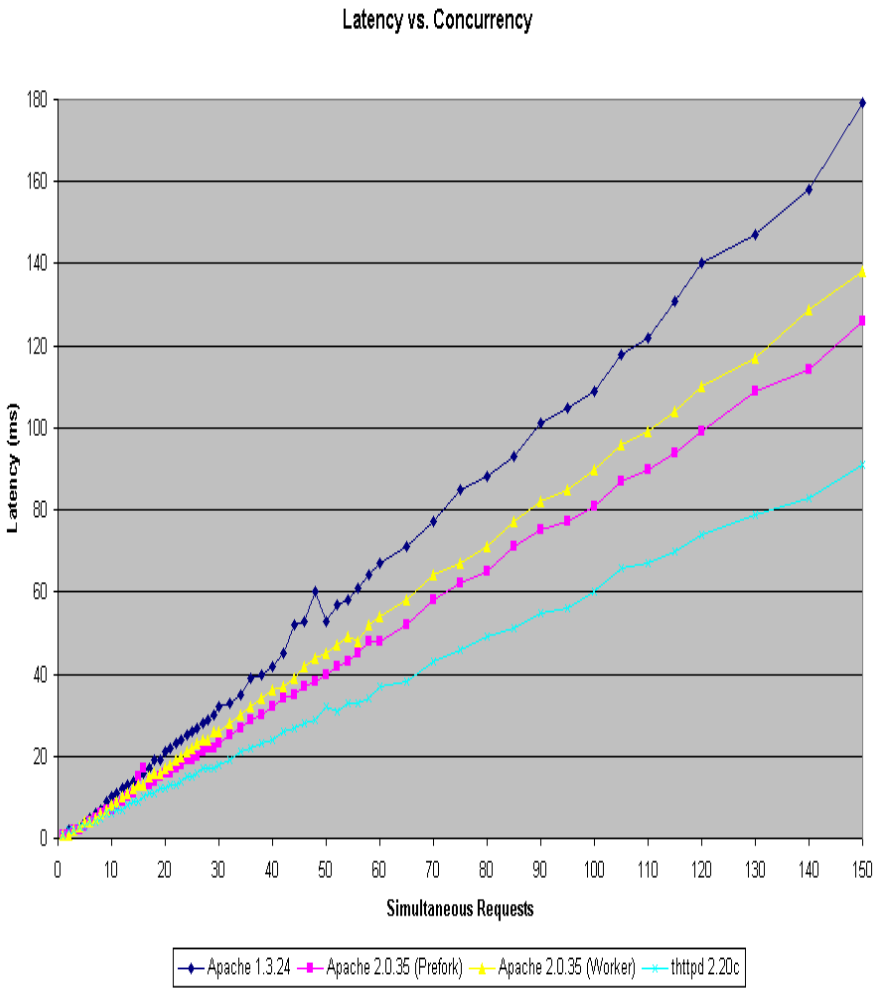
MEDIDAS BÁSICAS DEL RENDIMIENTO





EJEMPLO PARA UN SERVIDOR WEB





CARGA, EVALUACIÓN Y MEDIDA

Sistema real

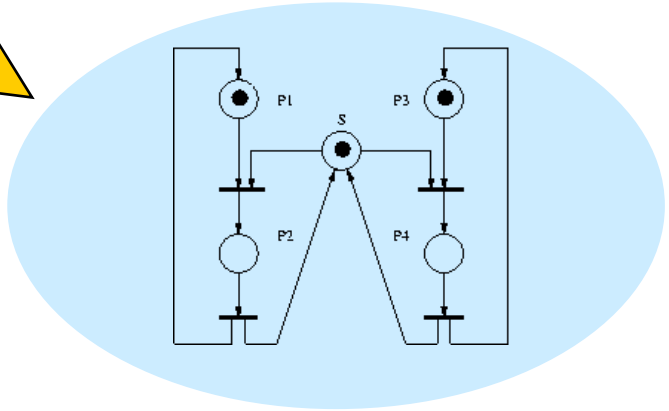
Carga real

Modelo de la carga real

Índices de rendimiento

Índices de rendimiento

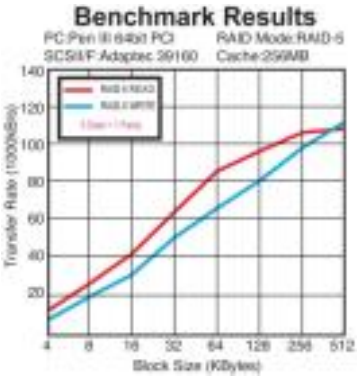
Modelo del sistema real



¿Son iguales?

Índices de rendimiento

¿DE QUÉ TÉCNICAS DISPONEMOS?



Métodos y herramientas para estimar los índices de prestaciones

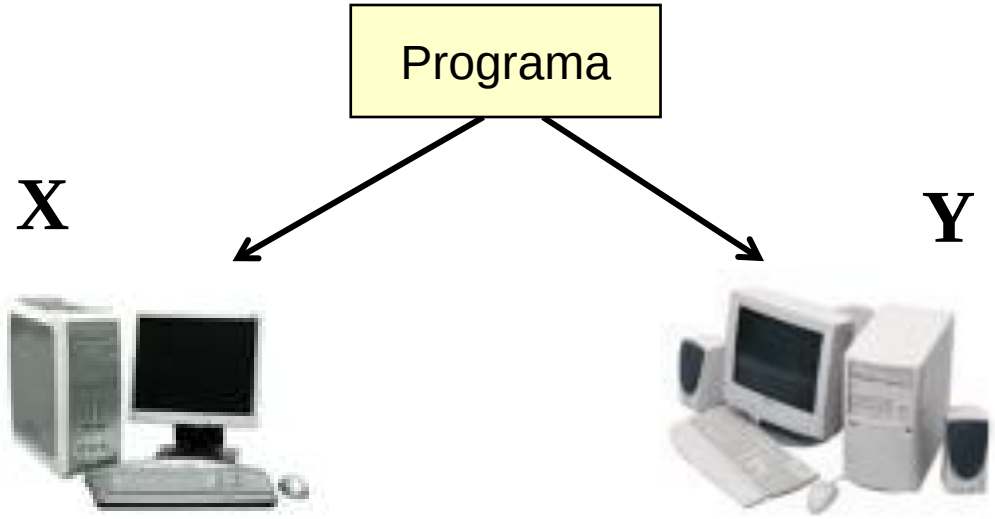
- Monitorización del sistema real
  - Herramientas de medida sobre el sistema real
- Referenciación (*benchmarking*) con sistemas reales o modelados
  - Comparación del rendimiento de sistemas
- Modelado
  - Reproducción del comportamiento del sistema
    - Métodos analíticos (redes de colas, cadenas de Markov, redes de Petri, ...)
    - Simulación discreta (CSIM, SMPL, Simula, ...)

COMPARACIÓN DE PRESTACIONES

Perspectiva actual

- Ejecutar los programas reales (o los más precidos a los programas reales) para evaluar el rendimiento de un sistema

El computador más rápido es aquel que ejecuta la aplicación en el tiempo más corto



COMPARACIÓN DE PRESTACIONES

Si quisiéramos calcular cuántas veces es más rápido el ordenador *A* que el *B*, para la ejecución de ese programa, bastaría dividir y obtener la aceleración (*speedup*) de *A* sobre *B*.

La aceleración representa el incremento de rendimiento de un ordenador respecto del otro y posee unidades, es decir, que en el ejemplo el ordenador *A* es tantas veces más rápido que *B*.

$$Aceleración = \frac{T_B}{T_A}$$

Se puede expresar la aceleración en porcentaje y, por tanto, expresaríamos que el ordenador *A* es un *n* % más rápido que *B*, mediante la fórmula:

$$Aceleración = \frac{T_B}{T_A} = 1 + \frac{n}{100}$$

EJEMPLO DE COMPARACIÓN DE PRESTACIONES

Un programa se ejecuta en 36 s en el computador VERDE y en 45 s en el computador ROJO

$$A = \frac{\text{Tiempo}_{\text{ROJO}}}{\text{Tiempo}_{\text{VERDE}}} = \frac{45\,\text{s}}{36\,\text{s}} = 1.25 = 1.0 + \frac{25}{100}$$

El computador VERDE es 1.25 veces **más rápido** que el ROJO

El computador VERDE es un 25% **más rápido** que el ROJO

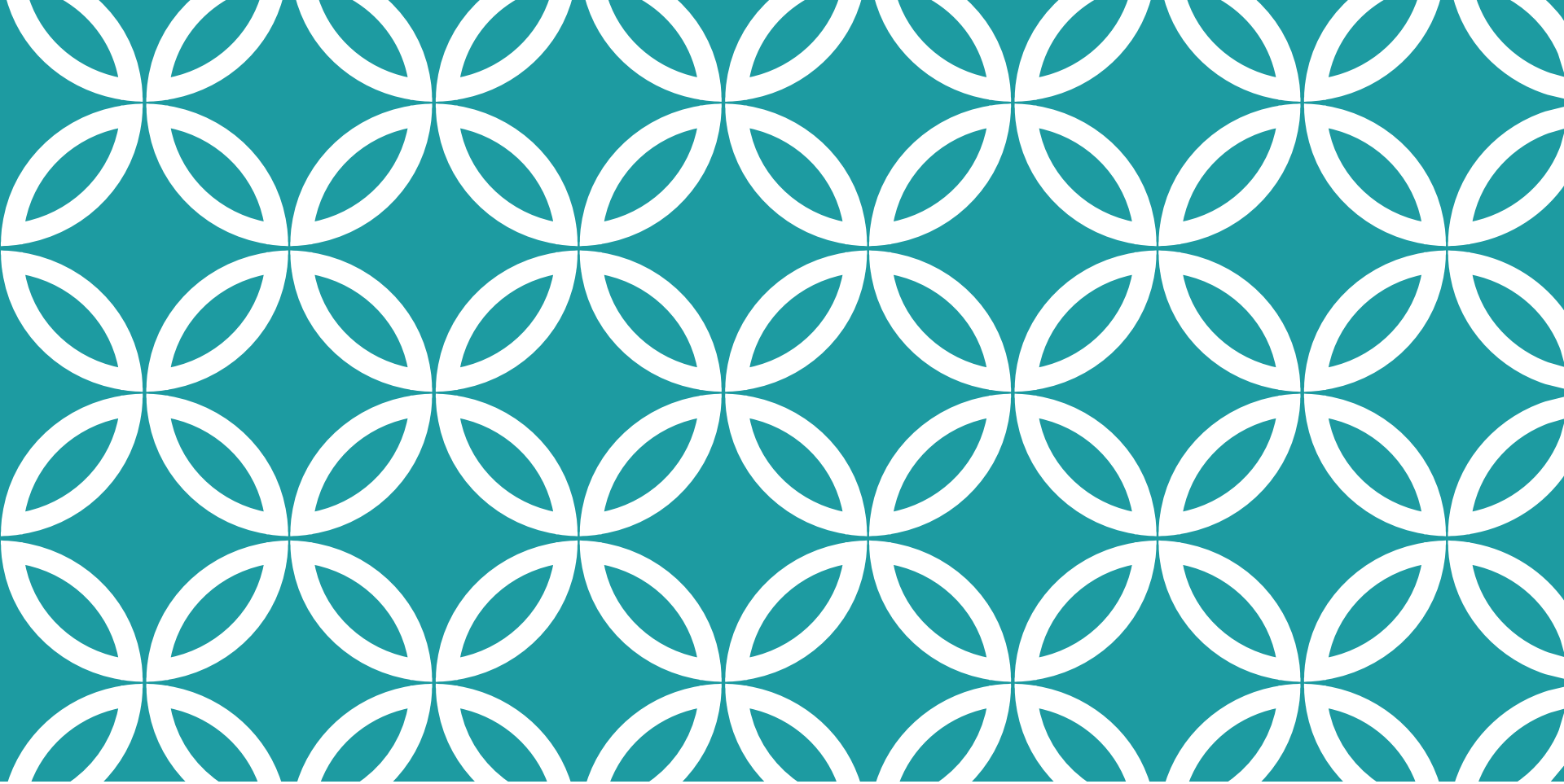
COMPARACIÓN DE PRESTACIONES

Por otro lado, se puede definir como cambio relativo del sistema *B* con respecto al sistema *A* como, para ver la diferencia sobre el tamaño del tiempo de ejecución:

$$Cambio\ Relativo_{B,A} = \frac{T_B - T_A}{T_A}$$

Si el cambio relativo es positivo, el sistema *B* es más lento que el sistema *A* y si el cambio relativo es negativo, el sistema *B* es más rápido que el sistema *A*.

- Ello es debido a que la referencia es el tiempo de ejecución del sistema A.
- Podemos ajustar la comparación para tener en cuenta el "tamaño" de las cantidades involucradas, definiendo, para valores positivos como el tiempo de ejecución



LÍMITES EN LA MEJORA DEL RENDIMIENTO

La ley de Amdahl
La ley de Gustafson
Ejemplos de aplicación

MEJORA DE UN SISTEMA

La mejora de un sistema no es ilimitada

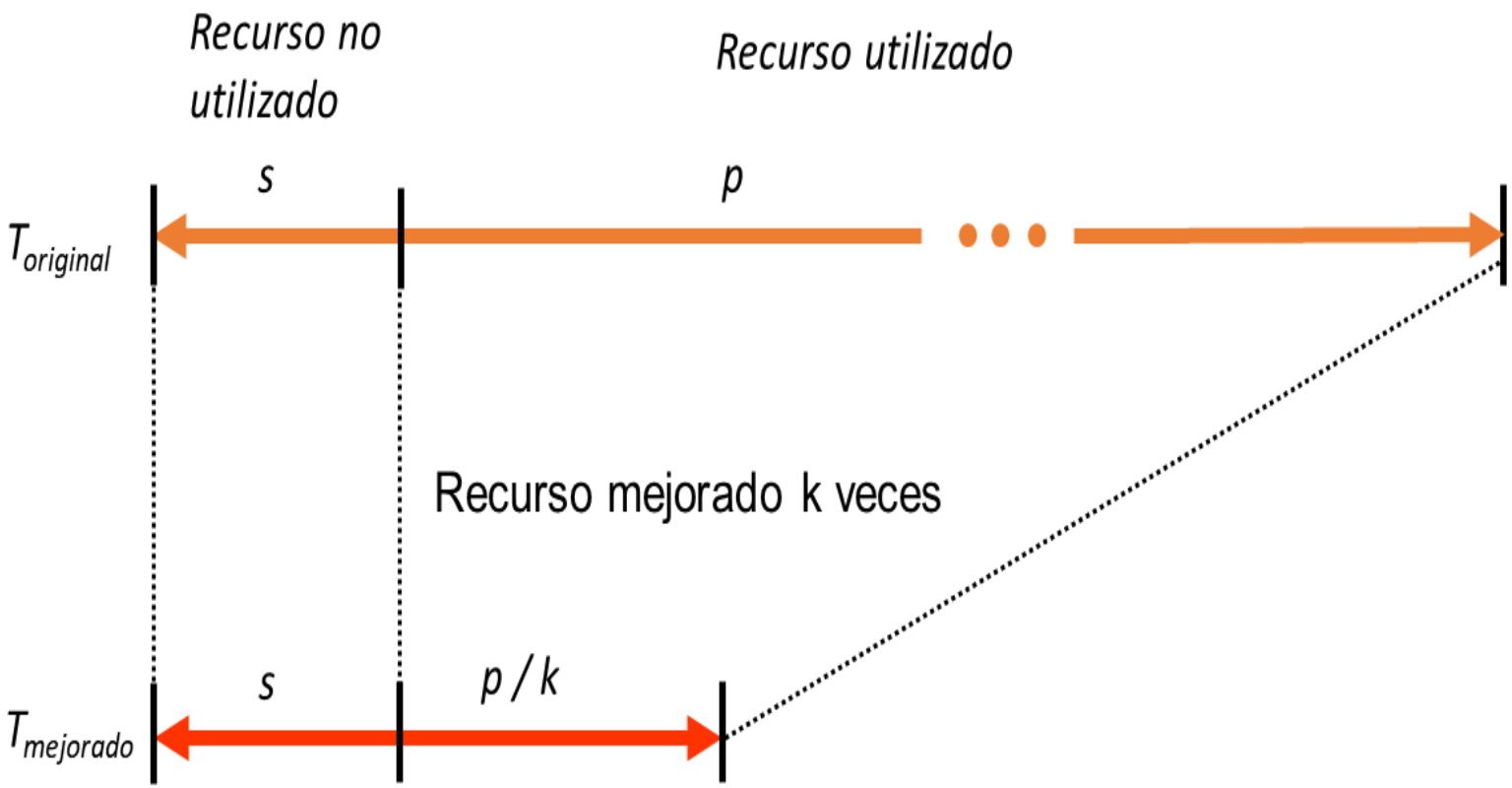
- Hay que saber hacia dónde dirigir los esfuerzos de optimización

La mejora de cualquier sistema debido a un componente más rápido depende del tiempo que éste se utilice

Discusión preliminar

- Un sistema tarda un tiempo $T_{original}$ en ejecutar un programa
- Mejoramos el sistema acelerando k veces uno de sus componentes
- Este componente se utiliza durante una fracción p del tiempo $T_{original}$
- ¿Cuál es la aceleración A del sistema global?

TIEMPO ORIGINAL VS. TIEMPO MEJORADO



TIEMPO ORIGINAL VS. TIEMPO MEJORADO

Si nos fijamos en la figura anterior, se puede expresar el $T_{original}$ en sus como la suma sus dos partes, puesto que $s + p = 1$:

$$T_{original} = T_{original} \cdot s + T_{original} \cdot p$$

Y entonces el $T_{mejorado}$ de forma análoga sería:

$$T_{mejorado} = T_{original} \cdot s + \frac{T_{original} \cdot p}{k}$$

De tal manera que la aceleración global se obtiene dividiendo ambos tiempos, tal como se definió anteriormente:

$$Aceleración = \frac{T_{original}}{T_{mejorado}} = \frac{s+p}{s+\frac{p}{k}} = \frac{1}{s+\frac{p}{k}} = \frac{1}{(1-p)+\frac{p}{k}}$$

$$(Ley\ de\ Amdahl)$$

LEY DE AMDAHL (1967)

¿Cuál es la aceleración A (*speedup*) del sistema completo después de acelerar k veces un componente?

$$Aceleración = \frac{T_{original}}{T_{mejorado}} = \frac{s+p}{s+\frac{p}{k}} = \frac{1}{s+\frac{p}{k}} = \frac{1}{(1-p)+(\frac{p}{k})}$$

Casos particulares de la ley

- Si $p = 0 \Rightarrow A = 1$: no hay ninguna mejora en el sistema
- Si $p = 1 \Rightarrow A = k$: el sistema mejora igual que el componente

(Notación: en muchos textos *f=p y 1-f=s*)

EJEMPLO DE CÁLCULO

La utilización de un procesador es del 60%

¿En cuánto aumentará el rendimiento del sistema si se duplica la velocidad del procesador ($k=2$)?

$$Aceleración = \frac{1}{1-p+\frac{p}{k}} = \frac{1}{(1-0{,}6)+(\frac{0{,}6}{2})} = 1{,}43$$

El rendimiento aumenta 1,43 veces
El rendimiento aumenta un 43%

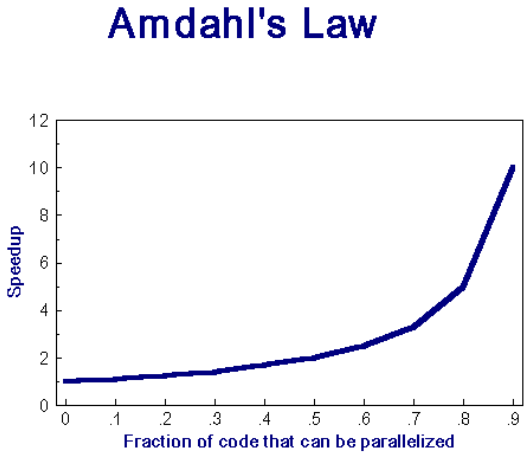
$$\lim_{k\to\infty} Aceleración = \lim_{k\to\infty} \frac{1}{1-p+\frac{p}{k}} = \frac{1}{1-p} = \frac{1}{s} = \frac{1}{0{,}4} = 2{,}5$$

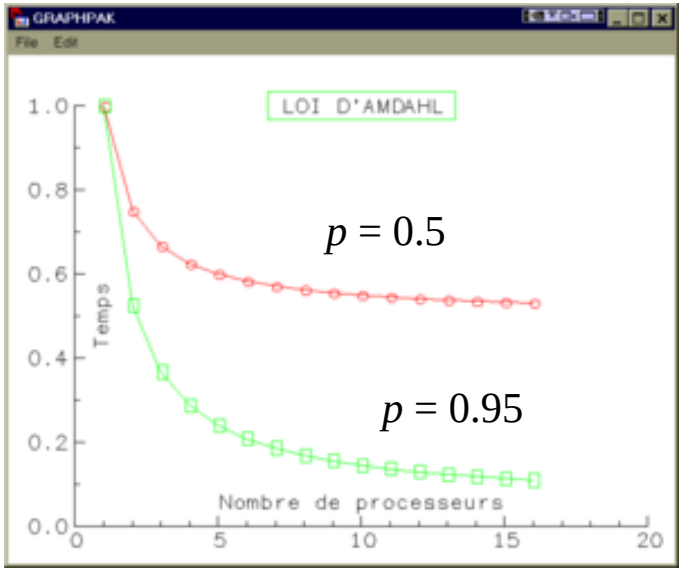
Aceleración máxima que se puede conseguir

CONTEXTO DE LA LEY DE AMDAHL

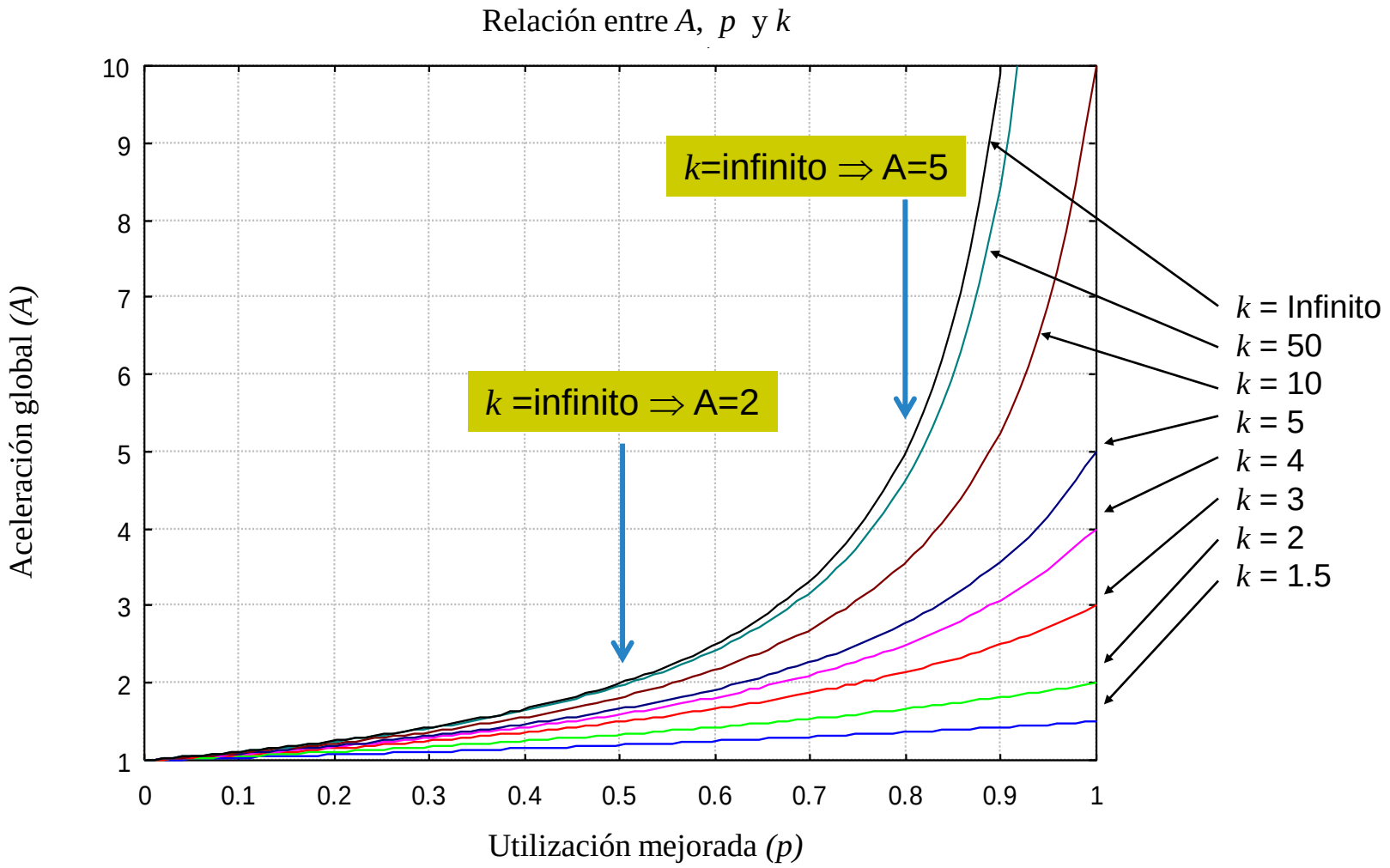
Artículo firmado por Gene Amdahl

- "*Validity of the Single Processor Approach to Achieving Large-Scale Computing Capabilities*", AFIPS Conference Proceedings, (30), pp. 483-485, 1967.
- Se utiliza para poner de manifiesto las limitaciones de los multiprocesadores en el cómputo paralelo





ANÁLISIS: RELACIÓN ENTRE A, p Y k



GENERALIZACIÓN DE LA LEY DE AMDAHL

Caso con una mejora solamente

$$Aceleración = \frac{T_{original}}{T_{mejorado}} = \frac{s+p}{s+\frac{p}{k}} = \frac{1}{s+\frac{p}{k}} = \frac{1}{1-p+\frac{p}{k}} \qquad (Ley\ de\ Amdahl)$$

Caso general con n mejoras

$$Aceleración = \frac{1}{(1-\sum_{i=1}^{n} p_i)+(\sum_{i=1}^{n} \frac{p_i}{k_i})} \qquad (Ley\ de\ Amdahl\ General)$$

LÍMITE DE LA LEY DE AMDAHL

La ley de Amdahl es realmente una consideración del límite de prestaciones bastante pesimista, fruto de los tiempos en que se enunció donde no había paralelismo, ni máquinas virtuales, ni *grid computing*, etc.

- Fijémonos que el límite del factor de mejora de un sistema mediante la optimización de un recurso viene dado precisamente por la fracción de tiempo cuando no se emplea ese recurso.
- Esto puede expresarse mediante el límite de la aceleración cuando el factor de mejora *k* es infinito (lo cual es imposible).

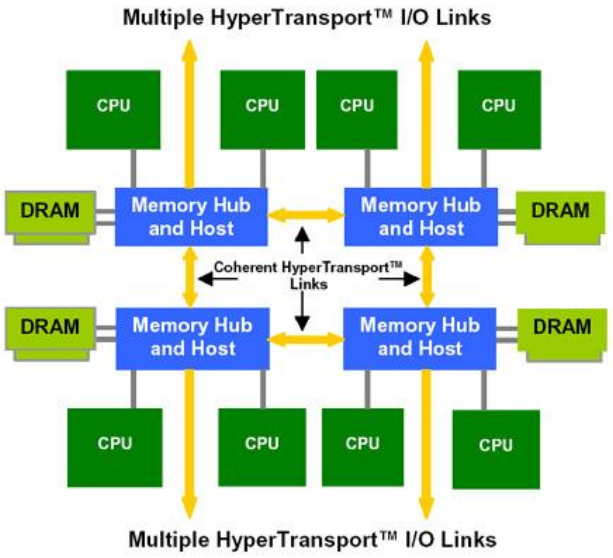
$$\lim_{k\to\infty} Aceleración = \lim_{k\to\infty} \frac{1}{1-p+\frac{p}{k}} = \frac{1}{1-p} = \frac{1}{s}$$

RENDIMIENTO DE MULTIPROCESADORES

Sistema con k procesadores

Aplicación a paralelizar

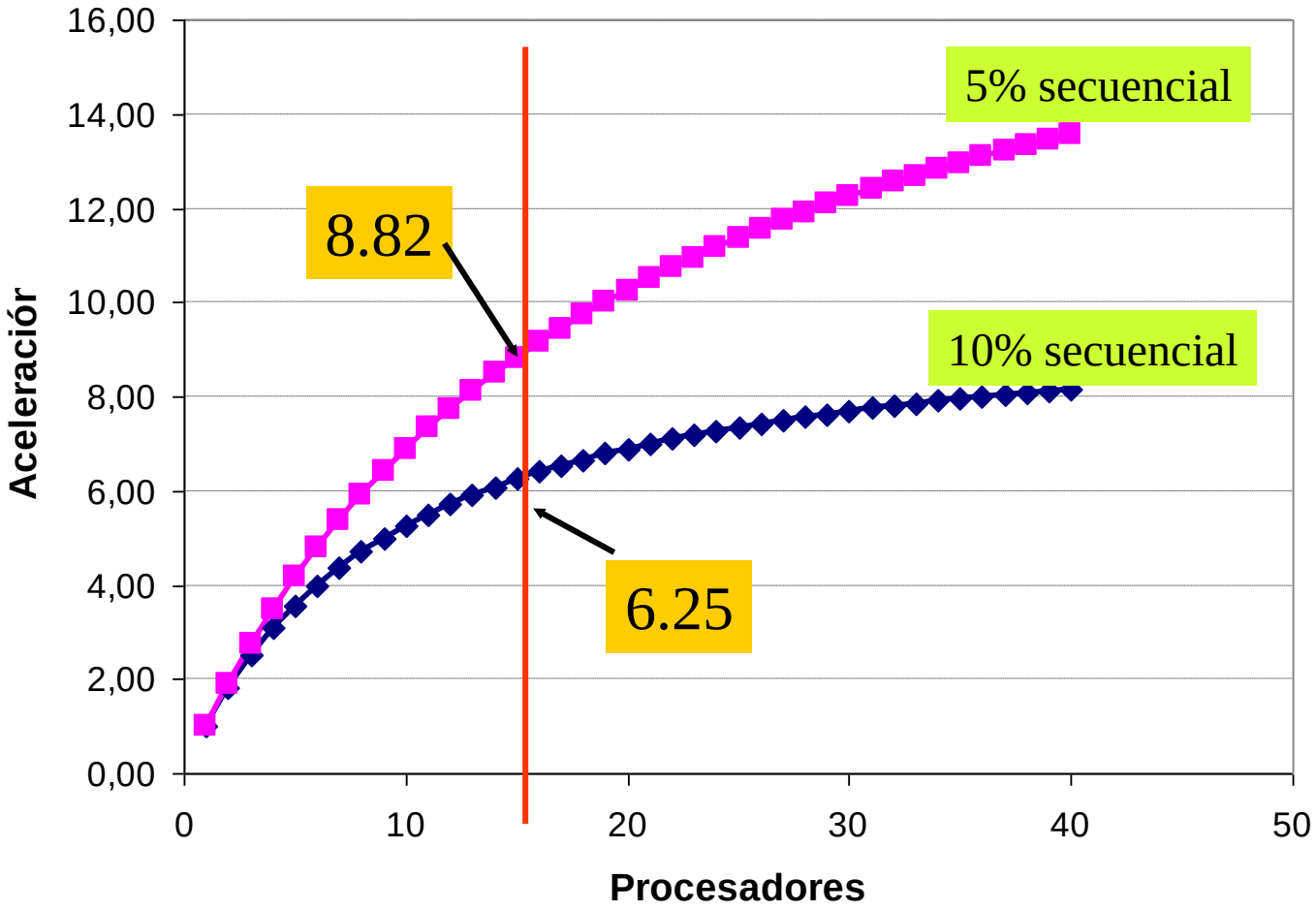
- Fracción secuencial = 0.10
- Fracción paralelizable = 0.90
  - ¿Cuál es la aceleración obtenida con 15 procesadores?
  - ¿Cuál es la aceleración máxima obtenible?



$$Aceleración = \frac{1}{s+\frac{p}{k}} = \frac{1}{0{,}1+(\frac{0{,}9}{15})} = 6{,}25$$

$$\lim_{k\to\infty} Aceleración = \lim_{k\to\infty} \frac{1}{1-p+\frac{p}{k}} = \frac{1}{1-p} = \frac{1}{s} = \frac{1}{0{,}1} = 10$$

EVOLUCIÓN DE LA ACELERACIÓN

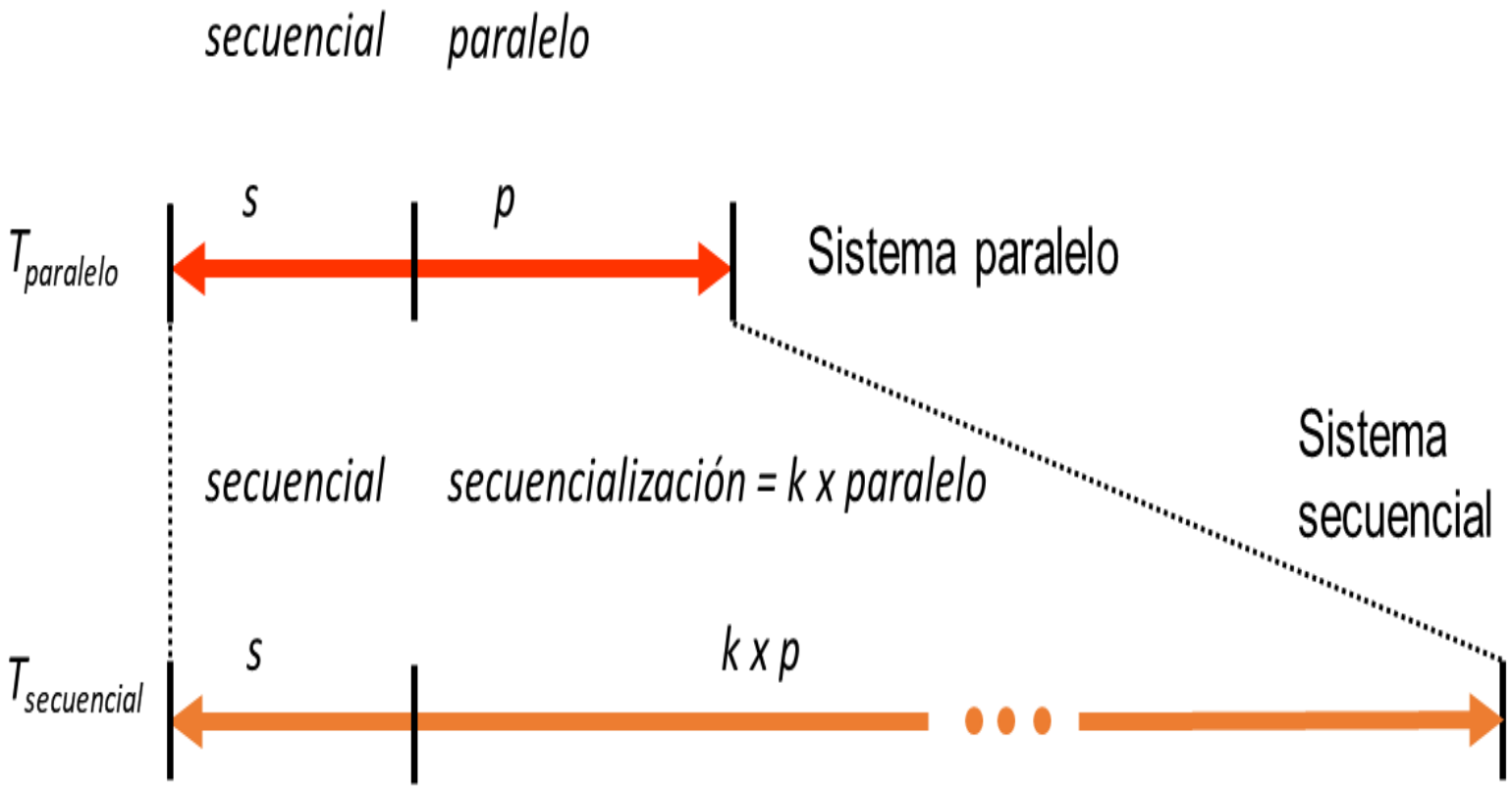


PLANTEAMIENTO DE GUSTAFSON

Amdahl enfatiza el aspecto más negativo del procesamiento paralelo

- Las máquinas paralelas se usan para resolver grandes problemas (meteorología, biología molecular…)
- Un computador secuencial nunca podría ejecutar un gran programa paralelo
- La cantidad de trabajo que se puede hacer en paralelo varía linealmente con el número de procesadores
- Con más procesadores se pueden acometer problemas de mayor coste computacional

LA ACELERACIÓN PROPORCIONAL



DERIVACIÓN DE LA LEY DE GUSTAFSON

Según Gustafson (ver figura anterior) siendo el tiempo de ejecución $T_{paralelo}$ para el procesamiento de un programa paralelo, que tiene una fracción secuencial s, al hacerlo totalmente secuencial su tiempo de ejecución $T_{secuencial}$ crecería linealmente en la parte paralela tantas veces como procesadores (k) tenga la máquina paralela.

- Es decir, lo ralentizaría linealmente en esa parte.

$$Aceleración = \frac{T_{secuencial}}{T_{paralelo}} = \frac{s+p\cdot k}{s+p} = \frac{s+p\cdot k}{1} = s + p \cdot k$$

DERIVACIÓN DE LA LEY DE GUSTAFSON

Como $s + p = 1$, simplificando y operando resulta la ley de Gustafson:

$$Aceleración = s + p \cdot k = s + (1 - s) \cdot k = k - s \cdot (k - 1)$$

Por consiguiente, si s es un valor pequeño entonces p es una fracción grande, con lo que la aceleración es aproximadamente k.

DERIVACIÓN DE LA LEY DE GUSTAFSON

Supongamos que el centro de datos con 20 servidores que se pueden usar en paralelo en un 90% por las aplicaciones de usuario ¿Cuál es la aceleración sobre un programa totalmente secuencial?:

$$Aceleración = \frac{T_{secuencial}}{T_{paralelo}} = k - s \cdot (k - 1) = 20 - 0{,}1\ (20 - 1) = 18{,}1$$

AMDAHL Y GUSTAFFSON

La realidad puede ser tan pesimista o más que el límite de la ley de Amdahl o tan optimista o más que la previsión de la ley de Gustafson. Ello depende del escenario tecnológico en el que estamos analizando el rendimiento.

Si vemos las dos leyes juntas, sus hipótesis son diametralmente opuestas. La ley de Amdahl modela la aceleración como de tamaño fijo, mientras que la ley de Gustafson modela la aceleración como de tamaño escalable.

$$Aceleración = \frac{1}{s+\frac{p}{k}} \qquad (Ley\ de\ Amdahl)$$

$$Aceleración = s + p \cdot k \qquad (Ley\ de\ Gustafson)$$

Para los extremos ambas fórmulas se comportan igual, $A = 1$ quiere decir que no hay mejora, con $s = 1$ (luego $p = 0$); mientras que $A = k$ cuando $s = 0$ (y $p = 1$).

AMDAHL Y GUSTAFFSON

La ley de Amdahl está hecha para analizar el efecto del paralelismo en la aceleración del sistema dado un problema de tamaño fijo.

- Establece que si una parte de una tarea, *p*, puede paralelizarse por un factor *k*, y el resto, *s*, no se puede paralelizar, entonces la porción que no se puede mejorar dominará rápidamente el rendimiento.

AMDAHL Y GUSTAFFSON

Para aplicaciones en las que el componente paralelo de la carga de trabajo se puede escalar linealmente mientras se mantiene el componente secuencial, Gustafson propuso como medida de aceleración esa escala.

- Esta suposición es válida para problemas de *grid computing* donde el programa a ejecutar se puede ajustar para aumentar la precisión de la computación.
- Por lo tanto, a medida que aumenta el número de procesadores, se aumenta la precisión.
- Esto aumenta la carga de trabajo paralela mientras mantiene la carga de trabajo secuencial.

METÁFORA DE LA CONDUCCIÓN

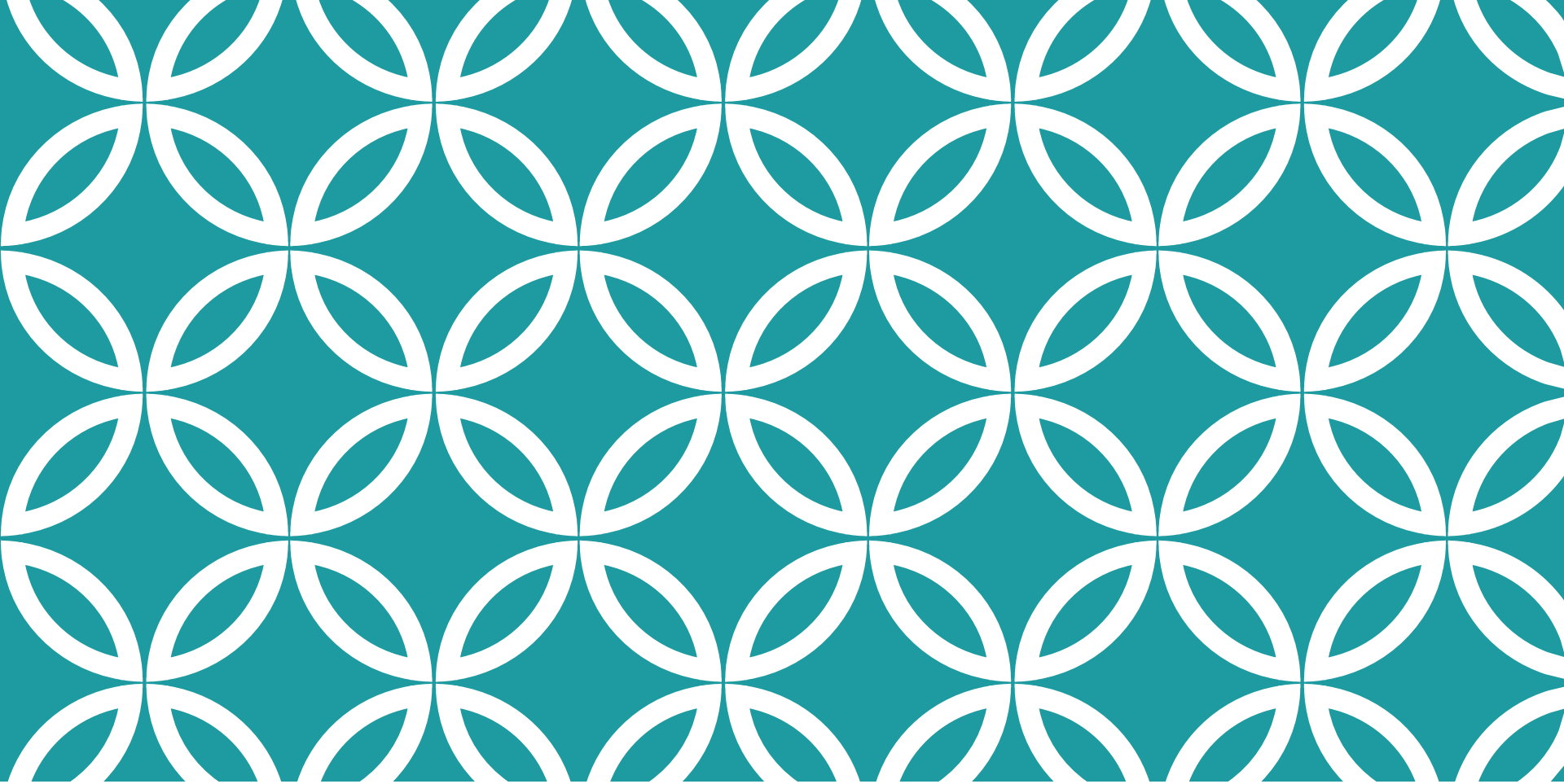
La ley de Amdahl aproximadamente sugiere:

- Supongamos que estás viajando en coche entre Palma y Porto Cristo (que se encuentran a 60 km de distancia) y ya has recorrido la mitad de la distancia, pero **has tardado una hora** debido al tráfico y las obras. Es decir tu promedio de velocidad es de 30 kph.
- Todavía te quedan 30 kms siguientes para llegar. Y ahora está despejado y puedes ir hasta 90 kph. Sin embargo, es imposible siquiera acercarse a esa velocidad promedio para llegar a Porto Cristo.
- De hecho, yendo a la velocidad de la luz a partir de ahora casi llegarías a 60 kph, debido a que ya llevas recorridos 30 kms en una hora!!!.

METÁFORA DE LA CONDUCCIÓN

La ley de Gustafson aproximadamente enuncia:

- Supongamos que has viajado en coche una hora a 30 kph. Si no te limitas a ir a Porto Cristo, es decir, **teniendo distancia y tiempo suficiente para viajar**, la velocidad promedio del coche podría alcanzar eventualmente los 90 kmph.
- Por ejemplo, si el coche tardó una hora viajando a 30 kmph, podría alcanzar las 90 kph de velocidad promedio conduciendo a 120 kph durante las siguientes 2 horas, o a 150 kph durante la siguiente hora, etc.



OTRAS MEDIDAS DE RENDIMIENTO

Eficiencia
Rendimiento/Coste
EDP

EFICIENCIA

Otra medida de prestaciones es la eficiencia, que se puede medir de varias formas.

Por ejemplo como la relación de aceleración y la cantidad de procesadores o unidades de procesamiento *k*.

- Se puede determinar el tiempo de respuesta de un programa secuencial y ejecutarlo en paralelo en *k* procesadores.
- Al calcular la aceleración del programa paralelo sobre el secuencial (o sobre ejecutarlo en un solo procesador), por ejemplo, se puede definir la eficiencia como la proporción (ratio) de la aceleración entre el número de procesadores..

$$Eficiencia\ (k) = \frac{Aceleración}{k}$$

EFICIENCIA

Otra formulación alternativa de eficiencia en una máquina puede ser la siguiente:

$$Eficiencia_w = \frac{Tiempo\ de\ trabajo\ efectivo}{Tiempo\ de\ trabajo\ total} = \frac{W}{W+Ov}$$

Es decir, considerar en el tiempo de trabajo que se pierde (*overhead*, en inglés), como una "desaceleración".

$$Eficiencia_w = \frac{W}{W+Ov} = \frac{1}{1+\frac{Ov}{W}}$$

EFICIENCIA (EJEMPLO)

Supongamos que el tiempo de respuesta de una tarea paralelizable en dos servidores idénticos es 9,27 ms y en 8 servidores paralelos es 1,30 ms. Ese misma tarea ejecutada en un servidor con dos máquinas virtuales tarda 19,28 ms y con 8 tarda 2,89 ms. Sus eficiencias son muy similares

$$Eficiencia_{virtual,2} = \frac{9{,}27}{19{,}28} = \frac{1}{1+\frac{10{,}01}{9{,}27}} = 0{,}48$$

$$Eficiencia_{virtual,8} = \frac{1{,}30}{2{,}89} = \frac{1}{1+\frac{1{,}59}{1{,}30}} = 0{,}45$$

Puesto que hemos tomado el tiempo de respuesta del servidor físico como patrón de tiempo efectivo, luego su eficiencia es 1 (y su aceleración):

$$Eficiencia_{físico} = \frac{9{,}27}{9{,}27} = \frac{1{,}30}{1{,}30} = 1$$

RENDIMIENTO Y OTRAS CARACTERÍSTICAS NO FUNCIONALES

El análisis de las prestaciones de un sistema, puede combinarse con otras magnitudes para combinar índices que son muy útiles para los administradores.

Si se comparase el rendimiento y el coste de dos ordenadores *A* y *B*, cuyos tiempos de respuesta al ejecutar un programa o tarea son T_A y T_B, y sus costes totales son C_A y C_B, respectivamente, se podría determinar cuál de los dos tiene una relación mejor en rendimiento/coste:

$$\frac{Rendimiento}{Coste} = \frac{1}{T \; x \; C}$$

EJEMPLO DE COMPARACIÓN DE COSTES

El computador VERDE cuesta 625 €

El computador ROJO cuesta 550 €

$$\Delta C = \frac{\text{Coste}_{\text{VERDE}}}{\text{Coste}_{\text{ROJO}}} = \frac{625\,€}{550\,€} = 1.14 = 1.0 + \frac{14}{100}$$

El computador VERDE es 1.14 veces **más caro** que el ROJO

El computador VERDE es un 14% **más caro** que el ROJO

EJEMPLO COMPARACIÓN PRESTACIONES/COSTE

En comparaciones de sistemas, idealmente, siempre interesará elegir aquellas opciones que maximicen el cociente prestaciones/coste. Recordemos que VERDE tardaba 36 s. y ROJO 45 s., respectivamente.

$$\frac{\text{Rendimiento}_{\text{ROJO}}}{\text{Coste}_{\text{ROJO}}} = \frac{1}{\text{Tiempo}_{\text{ROJO}} \times \text{Coste}_{\text{ROJO}}} = \frac{1}{45 \times 550} = 4.04 \times 10^{-5}$$

$$\frac{\text{Rendimiento}_{\text{VERDE}}}{\text{Coste}_{\text{VERDE}}} = \frac{1}{\text{Tiempo}_{\text{VERDE}} \times \text{Coste}_{\text{VERDE}}} = \frac{1}{36 \times 625} = 4.44 \times 10^{-5}$$

En este caso, el computador VERDE presenta una relación ligeramente más alta que el ROJO

EDP (ENERGY-DELAY PRODUCT)

Desde hace tiempo, los fabricantes de hardware, sobre todo los de procesadores y servidores, se han preocupado de producir dispositivos que maximicen el rendimiento minimizando el consumo energético.

En particular, la potencia eléctrica no es un buen indicador en procesadores y por extensión en los ordenadores que los alojan, puesto que depende de la frecuencia del reloj.

- Ello es debido a que al reducir la frecuencia del reloj, se reduce la potencia consumida, a costa del rendimiento.

EDP (ENERGY-DELAY PRODUCT)

Por tanto, parece interesante tener métricas que relacionen ambas. Por ejemplo, podríamos construir el siguiente índice que relaciona el tiempo de ejecución de instrucciones con su energía consumida, que se aplica a los procesadores, conocida como *EDP* (*Energy-Delay Product*).

$$EDP = Energía\ consumida \cdot Tiempo$$

EDP (ENERGY-DELAY PRODUCT)

Supongamos que $T_A = 3{,}0$ segundos y $T_B = 4{,}5$ segundos, el ordenador *A* es más rápido con una aceleración de $T_B / T_A = 4{,}5 / 3{,}0 = 1{,}5$. Supongamos que las potencias medias consumidas durante la ejecución del programa son $P_A = 100$ vatios y $P_B = 60$ vatios.

Con esos valores de potencia media y tiempo de respuesta podemos calcular la energía,
$E_A = 300$ vatios x segundo y $E_B = 270$ vatios x segundo.

Si calculamos el *EDP* de ambos, podemos decir que el ordenador *A* es más eficiente en energía y rendimiento que el ordenador *B* ya que:

$$EDP_A = E_A \cdot T_A = 300\ x\ 3{,}0 = 900\ Ws^2$$

$$EDP_B = E_B \cdot T_B = 270\ x\ 4{,}5 = 1.215\ Ws^2$$

ALGUNAS REFLEXIONES FINALES

Una mejora es más efectiva cuanto más grande es la fracción de tiempo en que ésta se aplica.

Para mejorar un sistema complejo hay que optimizar los elementos que se utilicen durante la mayor parte del tiempo (caso más común).

No sólo se debe observar la aceleración o la eficiencia en el rendimiento de los sistemas, sin tener en cuenta los tiempos de respuesta.

- Recordemos que el tiempo de respuesta es la medida original y más fiable del rendimiento

Se pueden establecer métricas tanto sobre costes o energía como de otras características no funcionales de los sistemas.

Lo importante es saber qué se quiere medir, es decir las unidades en las que se expresan los valores, cómo se pueden comparar y si se esperan a maximizar o minimizar los mismos.

DESARROLLAR LA FORMULACIÓN

ACTIVIDAD TEMA 1 (OBLIGATORIA)

Supongamos que en vez de usar un servidor para ejecutar un programa durante un tiempo t_o, se usan en paralelo k servidores durante un % de ese tiempo. El resultado es que mejora el rendimiento del sistema, ya que el programa ahora se ejecuta en un tiempo $t_m < t_o$.

¿Cómo calcularías ese % de tiempo?

Pista: cabecera de la transparencia…