

# Medición y análisis del rendimiento de un servidor

## Práctica 3

---

### Objetivo

La práctica está diseñada para que el alumno afiance los conocimientos teóricos relativos a la evaluación de sistemas informáticos, usando como ejemplo un sistema concreto: un servidor de información. Los aspectos más importantes abordados en la práctica son:

- 1) La caracterización (mediante la definición o especificación) de la carga que debe soportar un servidor y su inyección en el servidor usando herramientas de generación e inyección de carga.
- 2) La medición del funcionamiento del servidor usando monitores incorporados en el propio sistema operativo y en el inyector de carga.
- 3) La visualización y análisis de los datos medidos con el objeto de extraer un conjunto de métricas sobre el funcionamiento del servidor que permitan responder a preguntas como las siguientes:
  - ¿Cuál es la calidad de servicio (tiempo de respuesta) que puede ofrecer el servidor a los usuarios?
  - ¿Cuántos usuarios puede soportar el servidor simultáneamente para una determinada calidad del servicio?
  - ¿Cuál es la máxima productividad que se puede obtener del servidor?
  - ¿Se usan correctamente los recursos del servidor? ¿Cuándo se saturan?
  - ¿Cómo afecta el volumen de información manejada por el servidor a las otras métricas? Etc.

**Es importante guardar todas las mediciones ya que se utilizarán en futuras prácticas.**

### 1. Introducción al análisis a realizar

Los responsables del sistema informático de una compañía deben instalar un nuevo servidor de información y desean que el acceso a la información por parte de los usuarios cumpla unos determinados requisitos de tiempo de respuesta y que el servidor sea capaz de proveer una productividad mínima.

Los responsables del sistema disponen de un computador personal, actualmente desocupado, que desean usar para poner en marcha este proyecto. Pero desean que la empresa de informática encargada del proyecto realice una evaluación de la calidad de servicio (medida como tiempo de respuesta) que puede ofrecerse a los usuarios del sistema de información cuando se implementa en el computador personal. Como se dispone del servidor a utilizar, la evaluación se realiza usando la técnica de medición del modo siguiente:

Seleccionar las características promedio de las peticiones (transacciones) realizadas en el servidor.

Seleccionar un valor promedio para el tiempo de reflexión de los usuarios.

Se va incrementando la intensidad de la carga que debe soportar el servidor. Para ello se incrementa el número de usuarios así: 1, 10, 20, 30, 40, 50... Si se producen saltos cualitativos en el comportamiento del servidor entre dos números de usuarios (por ejemplo, entre 40 y 50), experimentar con más valores (por ejemplo, con 45).

Para cada número de usuarios anotar los siguientes resultados:

1) Productividad del sistema servidor, medida en peticiones/segundo.

2) Tiempo medio de respuesta que perciben los usuarios cuando realizan peticiones. Esta información se debe complementar con la desviación típica, y los valores mínimo y máximo del tiempo de respuesta. Una caracterización más exacta consiste en representar un histograma de frecuencias. Si no se desea una caracterización tan exacta se pueden usar los percentiles más significativos. La información sobre el tiempo de respuesta se puede dar de forma global (para todas las peticiones) o bien por cada clase de petición si es que se consideran varias clases o tipos de peticiones.

3) Información sobre la utilización media de todos los recursos del servidor, obtenida con el monitor de rendimiento. El objetivo es ver si algún componente del servidor está saturado y es el responsable de una degradación de su comportamiento.

El resultado final de esta evaluación consiste en decir que:

El servidor soporta P peticiones/segundo de determinadas características promedio, con un tiempo de respuesta medio T, O BIEN QUE...

El servidor soporta N usuarios caracterizados por un tiempo de reflexión Z y que realizan peticiones de determinadas características promedio, con un tiempo de respuesta medio T.

Además se debe indicar la utilización de recursos alcanzada para suministrar los servicios indicados anteriormente.

El análisis o evaluación del sistema debería completarse estudiando la influencia de las posibles variaciones de los parámetros considerados fijos en este análisis.

¿Qué ocurre cuando se incrementa y se reduce el tiempo de reflexión en un 25%? Si se mantiene el mismo tiempo de respuesta medio, ¿se soporta un 25% menos o un 25% más de usuarios respectivamente? En otras palabras, ¿es lineal la relación entre el tiempo de reflexión y el número de usuarios soportados cuando se mantiene el tiempo de respuesta medio?

¿Cómo afectan las características de la petición o peticiones realizadas al tiempo de respuesta medio y la productividad del servidor? ¿Cómo afecta la configuración del computador y el sistema operativo al tiempo de respuesta medio y la productividad del servidor? Hay que realizar nuevos experimentos modificando los parámetros que caracterizan las peticiones.

## 2. Inyección de carga y medición de funcionamiento

El primer paso a realizar antes de evaluar el funcionamiento de un servidor consiste en instalar en el servidor la aplicación que sirve a las peticiones que se realicen al propio servidor. En esta práctica se utiliza la aplicación sintética **ssif**, cuyo objetivo es emular de forma muy simplificada a un servidor de bases de datos, un servidor web, la combinación de ambos, etc.

Para servir cada petición o transacción, la aplicación **ssif** consume CPU, lee y escribe información de/en disco y usa memoria. Los valores medios de que caracterizan el consumo de recursos de cada transacción se le indican a la aplicación **ssif** en la línea de comandos cuando se la arranca. Cada equipo de trabajo (2 alumnos) debe utilizar unos valores medios concretos para la realización de la práctica.

En la tabla de esta página se indican en las primeras cuatro columnas los valores medios que deben utilizar los equipos de trabajo para cada grupo de prácticas de la asignatura. Así, por ejemplo, en la primera línea de la tabla se comprueba como los valores CPU=5000, Lect=5, Escr=5 y Mem=200 deben ser utilizados por el equipo de trabajo número 1 del grupo 1 de prácticas. Usar un **tiempo de reflexión** entre peticiones de **7 segundos** y una **razón de visitas** (quinto parámetro de **ssif**) de **4 para todos los grupos**.

Durante el período de tiempo que el servidor está recibiendo peticiones es necesario medir la utilización de los recursos del servidor, tales como tiempo de CPU, uso de los discos y la memoria, etc. En el S.O. Windows 2000 se dispone del monitor de rendimiento para realizar estas mediciones.

CPU	Lect	Escr	Mem	G-1	G-2	G-3	G-4	G-5
25000	10	5	200	1				1
25000	10	25	200		1			
25000	10	50	200			1		
25000	10	100	200				1	2
25000	200	5	300		2			
25000	200	25	300	2		2		
25000	200	50	300				2	
25000	200	100	300					3
25000	300	5	400			3		
25000	300	25	400	3				
25000	300	50	400		3			4
25000	300	100	400				3	
100000	10	5	200		4			
100000	10	25	200			4		
100000	10	50	200	4				
100000	10	100	200				4	5
100000	200	5	300				5	6
100000	200	25	300		5	5		
100000	200	50	300	5				
100000	200	100	300			6		
100000	300	5	400		6		6	
100000	300	25	400	6		7		
100000	300	50	400		7			
100000	300	100	400					7
250000	10	5	200	7	8		7	
250000	10	25	200					8
250000	10	50	200			8		
250000	10	100	200	8				
250000	200	5	300		9		8	
250000	200	25	300	9		9		
250000	200	50	300					9
250000	200	100	300				9	
250000	300	5	400				10	
250000	300	25	400		10			
250000	300	50	400			10		10
250000	300	100	400	10				

### 3. Presentación de resultados (Medición servidor)

Entregar utilizando la carátula dejada en la página web de la asignatura. No incluir el enunciado de la práctica.

**TAREA 1:** Medición y análisis de comportamiento del servidor al incrementarse la carga que soporta (número de usuarios que realizan peticiones). Se mantienen fijos todos los otros parámetros de la carga. Obtener como mínimo:

- Evolución del tiempo de respuesta promedio (seg) y el percentil-90 (seg).
- Productividad promedio del servidor (peticiones/seg).
- Utilización media de los recursos del servidor: %CPU, %Memoria, %Disco y %Red. Escoger el índice que te parezca más adecuado para representar el % de uso de memoria. Medir también la longitud media de la cola de disco, ya que será necesaria en prácticas posteriores

Presentar esta información en forma de curvas como se indica en la página siguiente y responder a las preguntas que se indican relativas a los datos obtenidos.

Antes de realizar los experimentos de medición es necesario seleccionar el intervalo de arranque y el intervalo de medición para los experimentos. Explicar y documentar las pruebas realizadas para hacer la selección.

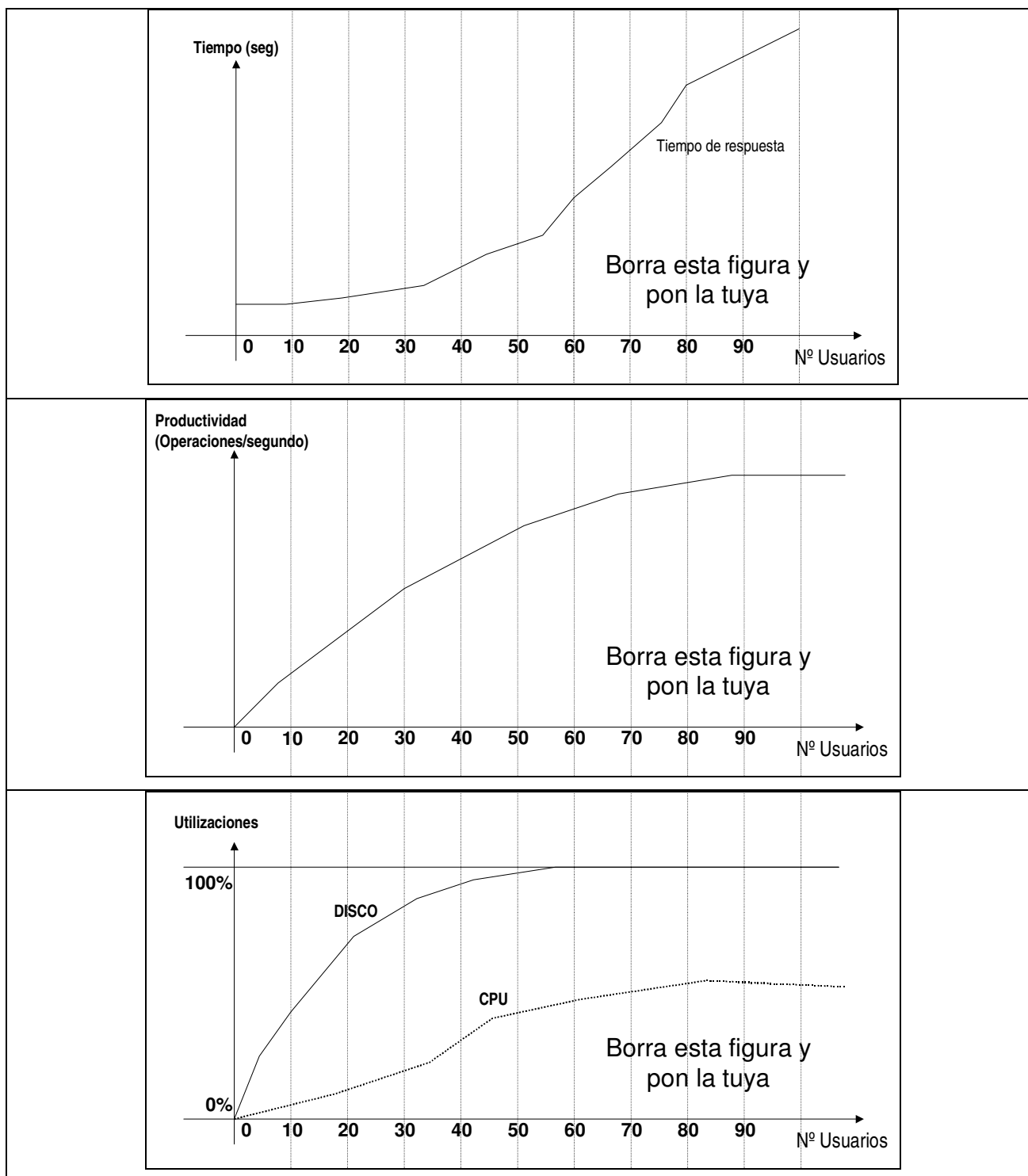
En función de las características de la transacción realizada por la aplicación ssif, para medir todos los regímenes de funcionamiento (productividad) del servidor (lineal, rodilla y saturación) el número máximo de usuarios a utilizar en los experimentos puede ser muy variable. Realizar experimentos para 1,2,3,4, ... usuarios resulta excesivo. Para empezar, se recomienda realizar una experiencia con un solo usuario. Esta experiencia permite determinar las características de la transacción realizada, esto es, el tiempo de respuesta de una transacción ejecutada en solitario, que en principio constituye el mínimo u óptimo que se puede esperar del sistema. A medida que se incremente el número de usuarios, este tiempo irá empeorando, aunque cuando hay unos pocos usuarios se pueden obtener tiempos medios mejores por aprovechamiento de cachés y del paralelismo en el *pipeline* del procesador. Posteriormente, para localizar los tres regímenes de funcionamiento es conveniente realizar experimentos para 20, 40, 60, 80,... usuarios, o bien 10, 20, 30, 40,... dependiendo del tiempo disponible. Una vez localizados los regímenes se realizan experimentos adicionales para generar más puntos en las curvas.

Para que cada punto representado en las curvas sea fiable sería necesario realizar varias réplicas de un mismo experimento. El número de muestras en cada réplica y el número de réplicas deberían ir aumentándose hasta obtener la precisión requerida, que debería contrastarse mediante intervalos de confianza. Debido al enorme trabajo que requeriría hacer las cosas bien, en esta tarea 1 se permite obtener cada punto de las curvas de una sola medición.

**CARACTERIZACIÓN DEL EXPERIMENTO DE MEDICIÓN**

Sistema:	CPU: P-IV ¿? Mhz ¿?MB cache / Ram: ¿?MB / Disco: IDE ¿?rpm / Red: Ethernet ¿? Mbps Sist Op: Win2000 / FichPag: ¿?MB
Carga:	Transacción: ¿?cpu ¿?lect ¿?escr ¿?ram 4 visitas / T reflexión: EXP(¿?) Intervalo arranque: ¿? segundos / Intervalo Med: ¿? Minutos

**RESULTADOS**



**CUESTIONES RELATIVAS AL ANÁLISIS DEL FUNCIONAMIENTO DEL SERVIDOR**

¿Cómo has calculado el % de uso de memoria?

¿Qué regímenes de trabajo puedes diferenciar en el funcionamiento del servidor en cuanto a la productividad según el número de usuarios y dónde están aproximadamente sus fronteras? ¿Se pueden apreciar claramente las fases de comportamiento lineal, rodilla de productividad y saturación?

¿Qué tiempo de respuesta medio se puede garantizar con el servidor de forma que los recursos del servidor no estén ni infrautilizados ni saturados? Usar como referencia una utilización del 90% del recurso que primero se satura.

Si se desea asegurar un tiempo medio de respuesta de 2 seg., ¿cuántos usuarios simultáneos soporta el servidor?

Si se desea asegurar que el 90% de las peticiones tengan un tiempo de respuesta inferior a 2 seg., ¿cuántos usuarios simultáneos soporta el servidor? Comenta la diferencia entre usar la media y el percentil 90 como métrica de la calidad del servicio que provee a los usuarios el servidor.

¿Cuál es la máxima productividad absoluta que se puede obtener de este servidor? ¿Son admisibles los tiempos de respuesta y las utilizaciones para la productividad máxima?

¿Los componentes del servidor (CPU, disco, etc.) trabajan equilibradamente, o sea, tienen niveles de utilización similares según se va incrementando la carga? ¿Cuál es el recurso que actúa como cuello de botella?

Si consideras necesario añadir alguna otra gráfica de algún otro contador o algún comentario para explicar el comportamiento del servidor, puedes y debes hacerlo.

**TAREA 2:** Se seleccionarán dos puntos de trabajo en la zona lineal, uno con pocos usuarios (al principio de la zona lineal) y otro, al que denominaremos punto nominal, y que estará situado al final de la zona lineal, pero sin entrar en la rodilla. El número de usuarios en cada punto dependerá de los parámetros con los que se hayan realizado los experimentos.

Para esos dos puntos, se llevará a cabo un análisis completo: realizar las réplicas necesarias y de la longitud adecuada para que el tiempo de respuesta en cada punto de trabajo se pueda expresar con al menos una precisión del 10% con un nivel de confianza del 95%. El análisis más serio en estos puntos se debe a que el primero servirá como base para obtener parámetros de funcionamiento del servidor y el segundo es el punto de trabajo nominal que se debería especificar al suministrar el servidor. Entregar:

1. El intervalo de confianza obtenido para el tiempo de respuesta en cada punto.
2. Una explicación de los pasos seguidos para obtenerlo. Recordar que se puede escoger utilizar la técnica de réplicas independientes o la de la media por lotes.
3. Intervalo de confianza obtenido y errores cometidos para la productividad con una confianza del 95% sólo en el punto nominal. Recordar que como este índice de prestaciones tendrá una varianza distinta de la del tiempo de respuesta, el error cometido con una confianza del 95% para la productividad será distinto del 10%.
4. Obtener el histograma de la variable tiempo de respuesta para los dos números de usuarios. Para construir el histograma es necesario que se haga una medición con un número de peticiones suficientemente grande, lo que se puede conseguir con experimentos largos, o varias réplicas. Como punto de partida considerar todas las peticiones realizadas para el cálculo del punto 1. Indicar los pasos seguidos para escoger los tamaños de celda en el histograma. **IMPORTANTE: Los histogramas deben construirse en frecuencias relativas**, es decir, el valor del eje de ordenadas debe estar comprendido entre 0 y 1, lo que se consigue dividiendo el número de peticiones que pertenecen a cada intervalo entre el número total de peticiones realizadas.
5. Para el punto de trabajo con pocos usuarios, ajustar la distribución del tiempo de respuesta. El objetivo de este ajuste será servir como valor de entrada en la práctica de simulación. Para llevar a cabo el ajuste, responder a estas preguntas:
  - 5.1. Atendiendo a los valores de la media y la mediana, y al de el coeficiente de asimetría, ¿es sesgada la distribución? Si lo es, ¿hacia la derecha o hacia la izquierda?
  - 5.2. ¿Sugiere el coeficiente de variación alguna familia de distribuciones?
  - 5.3. A partir del histograma y del análisis que has hecho en las dos preguntas anteriores, ¿podrías sugerir alguna familia de distribuciones?
  - 5.4. Para la familia de distribuciones considerada, calcula sus parámetros. Este paso se realiza con ayuda de las fotocopias de las distribuciones (Capítulo 6 del libro “Simulation, Modeling and Analysis”, Law and Kelton. Edt. McGraw-Hill). En el cuadro de cada distribución aparece el estimador de máxima verosimilitud (MLE en las tablas). Este valor constituye el parámetro o parámetros de la distribución estimada.
  - 5.5. El último paso consiste en realizar una comparación entre la distribución propuesta (teórica) y la real.
6. En caso de que no se ajuste a ningún tipo de distribución teórica, proponer al menos una distribución empírica.