

# Modelado analítico del rendimiento de un servidor

## Práctica 5b

---

### 1. Objetivo

En esta práctica el alumno debe combinar los conocimientos y los datos adquiridos en el bloque temático de *Medición* con los nuevos conocimientos sobre *Modelado de sistemas* para proponer, ajustar y validar modelos de comportamiento de un servidor de información.

En esta práctica se considerarán dos tipos de modelos:

(1) Modelos a nivel de sistema, también llamados modelos de entrada/salida. Son modelos que no tienen en cuenta la estructura interna del servidor y tan sólo tratan de explicar y reproducir las respuestas del sistema en función de las entradas aplicadas al mismo. A estos modelos de un sistema se les denomina de caja negra.

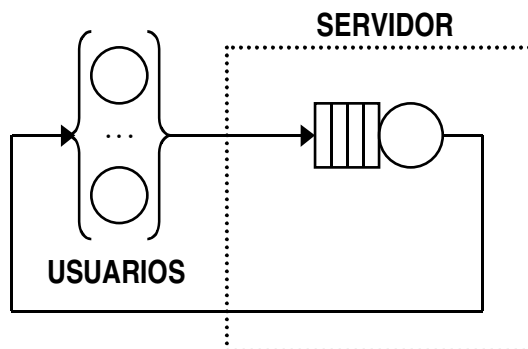
(2) Modelos a nivel de componente. En estos modelos se tienen en cuenta los diferentes elementos que componen el servidor, tales como la CPU, discos, etc. A estos modelos de un sistema se les denomina de caja gris.

Para ajustar y validar estos modelos se utilizarán los datos almacenados durante los experimentos de medición realizados en las prácticas previas.

En las prácticas siguientes, el alumno utilizará los modelos y la información obtenida a partir de ellos para configurar el servidor, por lo que se recomienda guardar cuidadosamente en disco toda la información manejada durante esta práctica.

## 2. Modelado a nivel de sistema

En esta sección se describen los pasos a realizar para ajustar los parámetros y validar un modelo de comportamiento de un servidor de información. El esquema del sistema (usuarios + servidor) se muestra en la figura siguiente.



Hay que comprobar que el prototipo del servidor que deseamos modelar no rechace peticiones de los clientes por falta de espacio en su cola de peticiones y que no aborte peticiones de los clientes por un exceso de tiempo (*time-out*) de procesamiento de la petición. Estas condiciones de funcionamiento real no pueden ser modeladas analíticamente y obligan a emplear técnicas de simulación. El servidor *ssif* verifica estas condiciones de funcionamiento.

El único parámetro que hay que ajustar en este modelo de una sola cola es la cadencia máxima de servicio ( $\mu$ ) de la cola, o lo que es lo mismo la cadencia máxima de servicio del servidor de información. En vez de usar el parámetro  $\mu$  se puede usar su inverso  $S=1/\mu$ , que es el tiempo de servicio del servidor. Por tanto el ajuste del modelo consiste en seleccionar el valor óptimo del parámetro  $S$ . A continuación se indican los pasos a seguir para parametrizar el modelo.

### PASO 1: SELECCIÓN DE PUNTOS DE FUNCIONAMIENTO

La precisión que se puede obtener de un modelo tan simple suele ser reducida. Disponiendo de un solo parámetro de ajuste es difícil que el comportamiento del modelo se ajuste al del sistema real en los tres regímenes clásicos de funcionamiento: lineal, rodilla y saturación. Por ello se debe elegir el punto o la zona de funcionamiento que el modelo debe ajustar preferentemente. También es posible usar todos los puntos medidos para la realización del ajuste, pero hay que tener en cuenta que si la productividad va disminuyendo en la zona de saturación en vez de mantenerse estable, el modelo no prevé este tipo de comportamiento, por lo que un ajuste basado en estas mediciones será muy deficiente.

Si el ajuste se basa en un solo punto se usará el punto de funcionamiento nominal. No obstante, no se recomienda usar las mediciones de un solo punto de funcionamiento para ajustar el modelo, sobre todo si no se han realizado las réplicas necesarias para que las variables que definen el punto (tiempo de respuesta y productividad) tengan el nivel de confianza necesario.

En la documentación de la práctica, presentar las gráficas de las curvas del tiempo de respuesta  $R$  y de la productividad  $X$ , resaltando los puntos seleccionados para realizar el ajuste del modelo.

## **PASO 2: SELECCIÓN DE LA DISTRIBUCION DEL TIEMPO DE SERVICIO**

Utilizar la distribución seleccionada en la práctica de medición para el experimento con 5 usuarios.

En caso de no disponer de información, suponer una distribución EXPonencial. En muchos casos deberá usarse obligatoriamente la distribución exponencial para que el modelo sea resoluble analíticamente.

## **PASO 3: SELECCIÓN DE UN RANGO DE VALORES DEL TIEMPO MEDIO DE SERVICIO PARA OBTENER EL VALOR ÓPTIMO**

Para obtener un valor inicial para el tiempo medio de servicio  $S$  se puede tomar el tiempo medio de respuesta obtenido cargando al servidor con un número de usuarios bajo (P.E. la medida de 5 usuarios). Una vez obtenido este valor del tiempo de servicio,  $S_{INI}$ , se elige un rango inicial de exploración de tiempos de servicio que puede ir de  $0.5 \cdot S_{INI}$  a  $1.5 \cdot S_{INI}$ .

También hay que elegir un incremento del tiempo de servicio para recorrer el rango que va de  $0.5 \cdot S_{INI}$  a  $1.5 \cdot S_{INI}$ . Por ejemplo si el tiempo medio de servicio es de 1 segundo, se recorre el rango de tiempos de servicio de 0.5 a 1.5 segundos en pasos de 0.02 para obtener 50 valores o en pasos de 0.01 para obtener 100 valores. Este bucle se programa en QNAP.

## **PASO 4: CÁLCULOS PARA CADA VALOR DEL TIEMPO DE SERVICIO**

Para cada valor del tiempo medio de servicio se resuelve el modelo para todos los puntos (número de usuarios concurrentes) de funcionamiento seleccionados. El experimento que se resuelve cada vez debe reflejar las mismas condiciones de funcionamiento del experimento de medición correspondiente (tiempo de reflexión medio y su distribución, número de usuarios, etc).

Calcular los errores de tiempo de respuesta y productividad para cada número de usuarios:

$ER = RP - RM$  El Error del tiempo de respuesta es la diferencia entre el tiempo de respuesta predicho por el modelo y el tiempo de respuesta medido en el prototipo.

$EX = XP - XM$  El Error de productividad es la diferencia entre la productividad predicha por el modelo y la productividad medida en el prototipo.

Calcular los errores medios de tiempo de respuesta y productividad para cada tiempo de servicio:

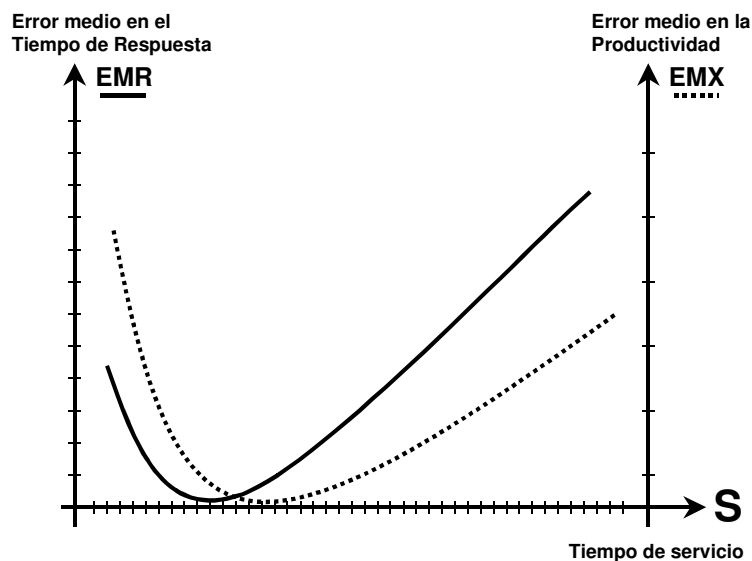
$EMR =$  Suma de una función de los errores de los tiempos de respuesta para todos los números de usuarios.

$EMX =$  Suma de una función de los errores de productividad para todos los números de usuarios.

Se pueden seleccionar diversas funciones de los errores antes de sumarlos. Por ejemplo se puede sumar los valores absolutos de los errores y dividir por el número de errores sumados. Otra opción es realizar la suma de los cuadrados de los errores y dividir por el número de errores sumados. Posteriormente se puede obtener la raíz cuadrada de esta magnitud o no. En general se puede usar cualquier función que combine los errores generados tras cada resolución del modelo en un solo error. Todos estos cálculos deben programarse en QNAP.

## **PASO 5: REPRESENTAR LOS ERRORES PARA TODOS LOS VALORES DEL TIEMPO DE SERVICIO SELECCIONADOS**

Las dos listas de errores medios de tiempo de respuesta y productividad para cada valor del tiempo de servicio seleccionado se representan en EXCEL. El comportamiento esperado de los errores medios es el siguiente:



Seleccionar el valor de S que proporciona un error medio mínimo. Si las dos curvas de error presentan el mínimo para el mismo valor de S, seleccionar ese valor. Si los mínimos se dan para valores de S distintos, hay que elegir entre:

- Minimizar el error de la productividad.
- Minimizar el error del tiempo de respuesta.
- Minimizar simultáneamente el conjunto de ambos errores.

La elección depende del uso que se le vaya a dar al modelo. En el caso de la práctica, seleccionar un valor de S que ajuste de modo aproximado el conjunto.

## **PASO 6: VALIDACIÓN DEL MODELO**

Una vez determinado el tiempo de servicio (distribución y valor medio) y conocidos los parámetros con los que se han realizado los experimentos de medición para un número creciente de usuarios, hay que realizar la validación del modelo en dos pasos:

1) Resolver el modelo de funcionamiento del servidor con las mismas condiciones en las que se han tomado TODOS los datos en la práctica de medición, usando el paquete QNAP.

2) Representar en la misma gráfica los dos comportamientos (el modelado y el medido) para cada métrica (tiempo de respuesta y productividad) y sus diferencias o errores. Comentar los resultados.

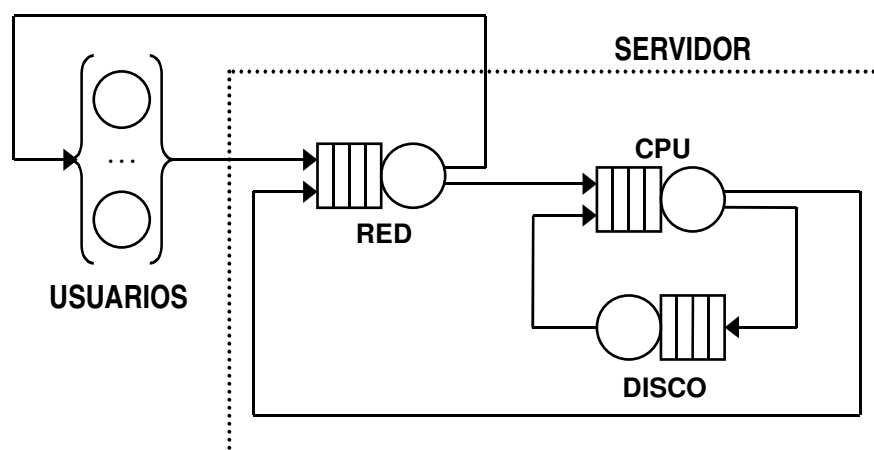
### 3. Modelado a nivel de componentes

El modelado a nivel de componentes más sencillo consiste en considerar al servidor compuesto por tres componentes: la CPU, el disco y la red. Este modelo básico podría ir complicándose al añadir otros componentes hardware/software o considerar aspectos de funcionamiento. Como elementos hardware/software a añadir estaría la disponibilidad de memoria. Como aspectos de funcionamiento se podría caracterizar el tipo de las peticiones considerando distintas clases en función de los recursos demandados por cada tipo de la petición.

El modelo debe predecir los valores de 5 variables básicas de funcionamiento: tiempo medio de respuesta (R), productividad (X), y las utilizaciones de los dispositivos básicos, Cpu ( $U_C$ ), Disco ( $U_D$ ) y Red ( $U_R$ ) con una precisión aceptable.

Los parámetros que permiten ajustar el funcionamiento del modelo son los tiempos (o cadencias máximas) promedio de servicio de las colas Cpu ( $S_C$ ), Disco ( $S_D$ ) y Red ( $S_R$ ) para las peticiones (sería mejor conocer incluso su distribución estadística para indicársela al programa QNAP) y la probabilidad de que una petición vaya al disco o a la red al terminar su servicio en la Cpu. Esta probabilidad esta directamente relacionada con la razón de visitas del Disco ( $V_D$ ).

Se utilizará el siguiente esquema de colas que se corresponde con el modelo más sencillo que se puede desarrollar del servidor a nivel de componentes:



El procedimiento de ajuste de los parámetros del modelo se basa en las mediciones realizadas y precisa de varios pasos.

#### PASO 1: OBTENCIÓN DE LAS DEMANDAS DE SERVICIO

Los modelos basados en redes de colas, para su resolución analítica, precisan que las demandas de servicio sean constantes cuando se incrementa el número de usuarios. Una primera comprobación del grado en el que el sistema se podrá modelar con una red de colas consistirá en representar las demandas totales de servicio de los tres componentes básicos del modelo:

$$D_C = U_C / X; \quad D_D = U_D / X; \quad D_R = U_R / X;$$

Las mediciones de las utilizaciones  $U_C$ ,  $U_D$  y  $U_R$  las proporciona el monitor del sistema operativo y la medición de la productividad del servidor la proporciona el inyector de peticiones.

Como las demandas de los componentes se expresan en unidades de tiempo se pueden representar en la misma gráfica que los tiempos de respuesta. Seleccionar los puntos (número de usuarios) para los que se desea ajustar el modelo. Para esos puntos las demandas de servicio de los componentes deberían ser aproximadamente constantes. Seleccionar los valores medios de las demandas para esos puntos.

En la documentación de la práctica, presentar las gráficas de las curvas del tiempo de respuesta  $R$  junto con las demandas  $D_i$ , la productividad  $X$  y las utilizaciones, resaltando los puntos seleccionados para realizar el ajuste del modelo.

## **PASO 2: DETERMINACIÓN DE LOS TIEMPOS MEDIOS DE SERVICIO Y LAS RAZONES DE VISITAS DE LOS COMPONENTES**

En cada componente se verifica la ley operacional de la demanda  $D_i = V_i \cdot S_i$ . Para descomponer la demanda de cada componente en  $V_i$  y  $S_i$  se hace lo siguiente:

Para la Red suponemos que por cada petición al servidor tenemos 2 visitas, una para la recepción de la petición y otra para el envío de la respuesta. Entonces  $V_R = 2$  y  $S_R = D_R / 2$ .

Para la Cpu y el Disco se puede utilizar la medida que da el monitor relativa a la productividad del disco (medida con el contador *Long. media de la cola de disco*)  $X_D$  y calcular  $V_D = X_D / X$ . En este modelo se verifica que  $V_C = 1 + V_D$ . Conociendo estas razones de visitas se pueden calcular directamente las probabilidades de transición para la cola Cpu para utilizarlas en un programa QNAP. Una vez estimadas las razones de visita se pueden calcular los tiempos de servicio directamente,  $S_C = D_C / V_C$  y  $S_D = D_D / V_D$ .

Para seleccionar la distribución a utilizar para los tiempos de servicio de los componentes no se dispone de medidas, por lo tanto suponer una distribución exponencial, que es la que se presupone por defecto en los modelos de colas.

## **PASO 3: OPTIMIZACIÓN DE LOS TIEMPOS MEDIOS DE SERVICIO DE LOS COMPONENTES**

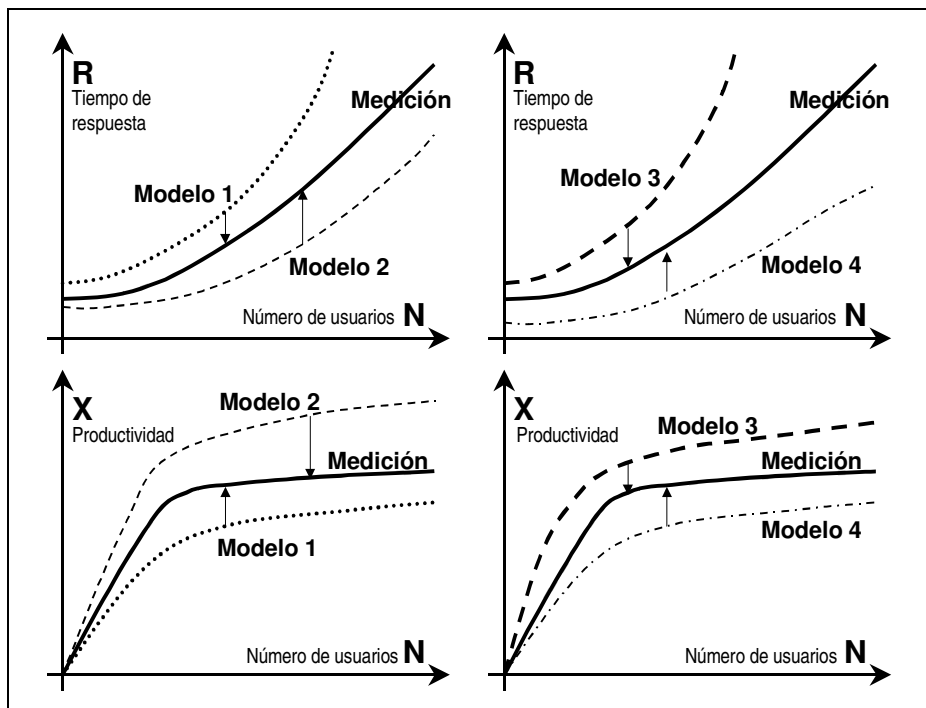
Hay que modificar 3 parámetros ( $S_C$ ,  $S_D$  y  $S_R$ ) para ajustar mejor la predicción de 5 variables ( $R$ ,  $X$ ,  $U_C$ ,  $U_D$  y  $U_R$ ). Es un problema de optimización complejo. Pero se puede plantear como una concatenación de tres problemas simples del modo siguiente:

- Con los tiempos medios de servicio estimados en el paso previo para los componentes se resuelve el modelo para todos los puntos (número de usuarios concurrentes) de funcionamiento seleccionados y se calculan los errores de tiempo de respuesta, productividad y utilizaciones para cada número de usuarios.
- Observar (representar si es preciso) los errores de las utilizaciones de los componentes. Se pueden calcular los errores medios de las utilizaciones. Si un componente presenta una utilización superior a la medida, reducir ligeramente su tiempo de servicio y viceversa.
- Realizar tanteos incrementando y/o disminuyendo el tiempo de servicio de los componentes hasta que las utilizaciones de los componentes que predice el modelo sean similares a las utilizaciones medidas con el monitor.

Antes de realizar el proceso de optimización manual propuesto conviene analizar si el ajuste de los tiempos de servicio puede mejorar o empeorar el ajuste de las productividades y/o los tiempos de respuesta. En general se verifica que una reducción de los tiempos de servicio de los componentes produce una reducción del tiempo de respuesta y un aumento de la productividad (y viceversa).

Por tanto, si por ejemplo se precisa reducir el tiempo de servicio de los componentes para ajustar sus utilizaciones, pero el tiempo de respuesta medio predicho por el modelo es también menor que el medido, debemos ser conscientes de que al reducir los tiempos de servicio de los componentes, el tiempo de respuesta que calcule el modelo será entonces mucho menor que el medido. En este caso, puede no tener sentido ajustar un poco más las utilizaciones de los componentes a costa de desajustar notablemente el tiempo de respuesta. Las mismas consideraciones que se han comentado sobre el tiempo de respuesta se pueden aplicar a la métrica de productividad.

La decisión de si merece la pena o no ajustar progresivamente los tiempos de servicio de las colas se ilustra con la figura siguiente, en la que las líneas continuas, rotuladas con la letra M, representan las mediciones del tiempo de respuesta R y de la productividad X.



En el modelo 1, los tiempos de servicio de las colas son excesivos, o lo que es lo mismo los componentes del modelo son más lentos que los reales. El modelo genera unos tiempos de respuesta mayores que los medidos y unas productividades menores que las medidas. En este caso una reducción gradual de los tiempos de servicio aproximará las dos curvas “1” a la curva M.

En el modelo 2, los tiempos de servicio de las colas son demasiado pequeños, o lo que es lo mismo los componentes del modelo son más rápidos que los reales. El modelo genera unos tiempos de respuesta menores que los medidos y unas productividades mayores que las medidas. En este caso un incremento gradual de los tiempos de servicio aproximará las dos curvas “2” a la curva M.

En los modelos 3 y 4 ambas curvas (R y X) están por encima o por debajo de las curvas medidas. Cambiando los tiempos de servicio de las colas no se puede reducir o aumentar simultáneamente las dos variables R y X, por lo que nunca será posible aproximar ambas curvas a las mediciones. El único ajuste posible en estos casos consiste en decidir cómo repartir el error entre el tiempo de respuesta y la productividad.

Teniendo en cuenta todas las consideraciones anteriores, explica la optimización que has realizado detallando todos los pasos y decisiones tomadas. Minimizar el volumen de texto aportado y maximizar los resultados de tipo gráfico.

#### **PASO 4: VALIDACIÓN DEL MODELO**

Una vez determinados los tiempos de servicio y conocidos los parámetros con los que se han realizado los experimentos de medición para un número creciente de usuarios, hay que realizar la validación del modelo en dos pasos:

- 1) Resolver el modelo de funcionamiento del servidor con las mismas condiciones en las que se han tomado TODOS los datos en la práctica de medición, usando el paquete QNAP.
- 2) Representar en la misma gráfica los dos comportamientos (el modelado y el medido) para cada métrica (tiempo de respuesta y productividad) y sus diferencias o errores. Comentar los resultados.

## **4. Presentación de resultados**

- El alumno debe construir un modelo “aceptable” para el servidor **a nivel de sistema** para las mismas condiciones de trabajo (tiempo de reflexión y distribución de peticiones) que las de la práctica de medición. Documentar todos los pasos realizados para la selección y ajuste de los parámetros del modelo, así como la validación del mismo.
- El alumno debe construir un modelo “aceptable” para el servidor **a nivel de componentes** para las mismas condiciones de trabajo que en el apartado anterior. Igualmente deben documentarse todos los pasos realizados para la selección y ajuste de los parámetros del modelo, así como la validación del mismo.
- EN AMBOS CASOS DEBE INCLUIRSE EL CODIGO QNAP DE CADA MODELO.

**Conservar los programas desarrollados en esta práctica, pues se hará uso de ellos en prácticas sucesivas.**