# Modelado analítico del rendimiento de un servidor
# Práctica 5

# 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.

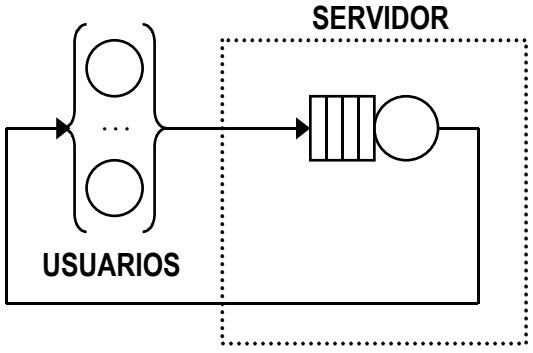
(2) Modelos a nivel de componente. En estos modelos se tienen en cuenta los diferentes elementos que componen el servidor, tales como la CPU, discos, etc. A estos modelos de un sistema se les denomina de caja gris.

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 programa *servidor_SCES* 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 (μ) de la cola, o lo que es lo mismo la cadencia máxima de servicio del servidor de información. En vez de usar el parámetro μ 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. Por ello, se recomienda utilizar tres puntos cercanos al punto de funcionamiento nominal.

**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.** Estos puntos definen tres ternas de valores conocidos (nº usuarios, tpo. respuesta medida y productividad medida): N1, RM1 y XM1; N2, RM2 y XM2; N3, RM3 y XM3.

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

Debería utilizarse la distribución obtenida a partir de la práctica de medición para el experimento con 5 usuarios. Sin embargo, por simplicidad y para hacer el modelo resoluble analíticamente, deberá usarse obligatoriamente la distribución exponencial.

## 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 (por ejemplo, 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 $0.1\times S_{INI}$ a $2.1\times S_{INI}$.

También hay que elegir un incremento del tiempo de servicio para recorrer el rango que va de $0.1\times S_{INI}$ a $2.1\times S_{INI}$. Por ejemplo si el tiempo medio de servicio es de 1 segundo, se recorre el rango de tiempos de servicio de 0.1 a 2.1 segundos en pasos de 0.1 para obtener 20 valores o en pasos de

0.05 para obtener 40 valores. Este barrido de valores se indicará en la pestaña "*What-if*" de la herramienta *JMVA* de la aplicación *JMT*.

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

La aplicación a utilizar tiene limitaciones a la hora de especificar variaciones de parámetros así como la posibilidad de exportar datos a ficheros. Se procederá de la siguiente forma:

1. Se creará con la herramienta *JSIMgraph* el modelo de red de colas que define el sistema a resolver. Sobre el modelo incluiremos todos los parámetros necesarios. Inicialmente se colocará en la pestaña de la sección de servicio del servidor el valor inicial del tiempo a explorar ($0.1 \times S_{INI}$).
2. Se resolverá el modelo en cada uno de los puntos de funcionamiento elegidos en el PASO 1. Para ello, desde *JSIMgraph* se invocará al resolutor analítico *JMVA*. En este tendremos que modificar:
   - En la pestaña *"Clases*" se inicializará con el número de usuarios del primer punto de funcionamiento (N1).
   - En la pestaña "*What-if analysis*" se elegirá como "*Control parameter*" la variación de las demandas de servicio, *Service Demands*. Nos aparecen nuevos campos a rellenar, en el desplegable *Station* se indicará que la cola a la que se aplica la variación de tiempos será la del servidor, en la caja "*To*" se introduce el tiempo final de exploración ($2.1 \times S_{INI}$) y en la caja "*Steps*" configuramos el número de valores intermedios que se desea.

   Una vez hecho esto pulsamos en *solve* y se obtendrán los valores del modelo[1]. Estos valores se pueden exportar a un fichero pulsando el botón *Export to file* dentro de la pestaña *Textual results*. El fichero se llamará *JMVAresults.tsv* y se generará en la misma carpeta donde esté el ejecutable de la aplicación JMVA.
3. En un fichero Excel, colocaríamos para cada número de usuarios de los puntos de funcionamiento sendas columnas: una encabezada con el tiempo de respuesta medido (TR1) y otra con la productividad medida (X1). En las sucesivas filas se copiarán desde el fichero *JMVAresults.tsv* los valores de tiempo de servicio probado, tiempo de respuesta obtenido y productividad para la cola del servidor (importar el archivo en Excel, cambiando puntos por comas si es necesario).

   Se repetirá el proceso para el número de usuarios N2 y N3. El número de usuarios se cambia en la pestaña "*Classes*" de la herramienta *JMVA*.
4. Calcular los errores de tiempo de respuesta y productividad para cada número de usuarios. Con los valores obtenidos con el modelo analítico, se calcula el error cometido en el cálculo del tiempo de respuesta y de productividad. Para ello se compara cada tiempo de respuesta y productividad obtenidos con los valores medidos que encabezaban las columnas en la Excel creada. Tendríamos así tres columnas de errores para los tiempos de respuesta y otros tantos para las productividades que nos permitirían calcular estos valores:

   **ERij = RPij – RMi**

   donde **i** es el número de usuarios en cada punto y **j** es en tiempo de servicio probado.

[1] Los valores de productividad son correctos, pero en el caso de los tiempos de respuesta hemos de fijarnos solo en el del servidor, pues el tiempo de respuesta del sistema incluye el tiempo de reflexión, que no debería considerarse.

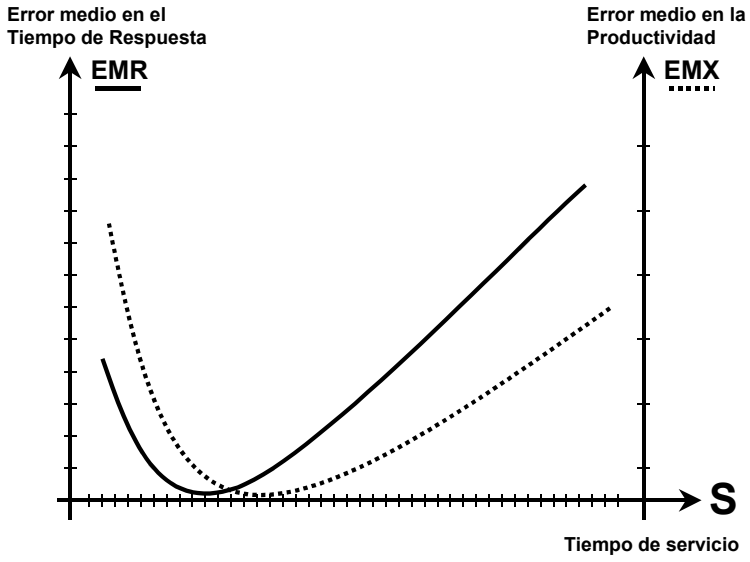
El error del tiempo de respuesta (ERij), es el valor de la diferencia entre el tiempo de respuesta predicho por el modelo cuando se resuelve para el tiempo de servicio j y el número de usuarios i (RPij) y el tiempo de respuesta medido en el prototipo con i usuarios (RMi).

**EXij = XPij – XMi** El error de productividad (EXij), es el valor de la diferencia entre la productividad predicha por el modelo cuando se resuelve para el tiempo de servicio j y el número de usuarios i (XPij) y la productividad medida en el prototipo con i usuarios (XMi).

5. Calcular los errores **absolutos** medios de tiempo de respuesta y productividad para cada tiempo de servicio. Se considera esta forma de resumir los errores por simplicidad. De esta forma para cada tiempo de servicio habremos calculado un valor del error cometido por el modelo analítico al estimar el tiempo de respuesta en los tres puntos de referencia. Del mismo modo habrá un valor que cuantifique el error cometido por el modelo analítico en el caso de la productividad.

## 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.

En cualquier caso, debe indicarse en la memoria de forma clara y precisa el valor de tiempo de servicio obtenido. Este valor se utilizará para resolver el modelo y proceder a su validación.

**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. Para ello puede hacerse uso de la pestaña "*What-if*" de la herramienta *JMVA* eligiendo como "*Control parameter*" el número de clientes. Al igual que para el tiempo de servicio se especificará el valor final, que coincidirá con el número de usuarios máximo que hallamos medido, y el número de iteraciones. Esta alternativa tiene el problema de que los valores intermedios tal vez no se ajusten a los que hemos medido. Si ese fuera el caso, asegurarse de que el rango de valores se mantiene y hay suficientes medidas intermedias. Alternativamente, se podría resolver el modelo de forma individual en los puntos en los que hemos realizado las mediciones, si bien esta solución puede resultar más laboriosa.

2) Representar los dos comportamientos (el modelado y el medido) en las gráficas de cada métrica (tiempo de respuesta y productividad). **Marcar los puntos medidos en cada línea**. Comentar los resultados.

De la observación de las gráficas podrá determinarse la calidad del ajuste del modelo, y si este ajuste es mejorable modificando el valor del tiempo de servicio utilizado. La operación de ajuste busca reducir la diferencia entre las gráficas medidas y las obtenidas por el modelo. **Si es posible mejorar el ajuste, indica cómo**.

En caso de haber realizado el ajuste, indicar de forma clara y precisa el tiempo de servicio finalmente obtenido.

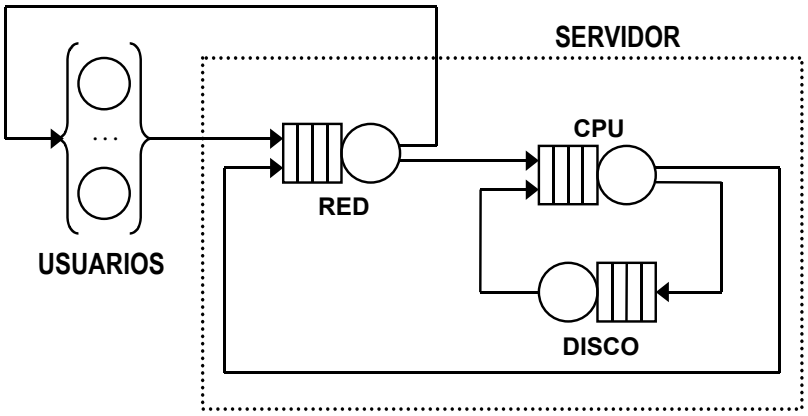
# 3. Modelado a nivel de componentes

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.

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 de resolución) y la probabilidad de que una petición vaya al disco o a la red al terminar su servicio en la CPU. Esta probabilidad está 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:

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

Ha de mantenerse la coherencia de unidades. **Las utilizaciones se expresan en tanto por uno**, y si la productividad viene en peticiones por segundo, la demanda se expresa en segundos.

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.

En caso de disponer de un dispositivo con varios servidores, como es el caso de la CPU, su demanda habrá que multiplicarla por el número de servidores (núcleos), es decir:

$D_C = (U_C \times N^o\text{núcleos}) / X$

Esta expression se obtiene a partir de la cola M/M/m con varios servidores. En dicho tipo de cola la utilización se obtiene a partir de la expresion: $U_i = X_i/(m\times \mu_i)$ donde $X_i$ es la productividad de la cola, m es el número de servidores y $\mu_i$ es la cadencia de servicio, es decir $1/ts_i$. Según la ley de flujo forzado $X_i = X \times V_i$, y además $D_i = V_i \times S_i$. Operando se llegaría a la expresion indicada. Posteriormente en el modelo de colas del sistema, la cola que represente a la CPU debería tener tantos servidores como núcleos disponga.

Para obtener las demandas representativas, se procede de la siguiente forma:

1. Representar las gráficas de evolución de la demanda obtenida para cada componente al variar el número de usuarios.
2. Seleccionar tres 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.
3. Seleccionar como demandas para los cálculos, los valores medios de las demandas para esos puntos en cada componente.

**En la documentación de la práctica, presentar las gráficas de las curvas del tiempo de respuesta R, las demandas Di, 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 $D_i=V_i \times S_i$. Para descomponer la demanda de cada componente en Vi y Si 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 (contador *Transferencias/s*) $X_D$ y calcular $V_D = X_D/X$. En este modelo se verifica que $V_C=1+V_D$. Una vez estimadas las razones de visita se pueden calcular los tiempos de servicio directamente: $S_C=D_C/V_C$ y $S_D=D_D/V_D$. Además, conociendo las razones de visitas se pueden calcular directamente las probabilidades de transición para la cola CPU para utilizarlas en el programa de resolución: $P_{CPU\text{-}Disco} = V_D/V_{CPU}$ y $P_{CPU\text{-}RED} = 1/V_{CPU}$.

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.

En la memoria debe presentarse una tabla con los valores calculados para las demandas, razones de visita, tiempos de servicio y probabilidades para cada componente.

## PASO 3: OPTIMIZACIÓN DE LOS TIEMPOS MEDIOS DE SERVICIO DE LOS COMPONENTES Y VALIDACIÓN DEL MODELO

El proceso de optimización es previo a la validación. La optimización consiste en modificar los parámetros ($S_C$, $S_D$ y $S_R$) para ajustar mejor la predicción de las variables (R, X, $U_C$, $U_D$ y $U_R$). Es un problema de optimización complejo.

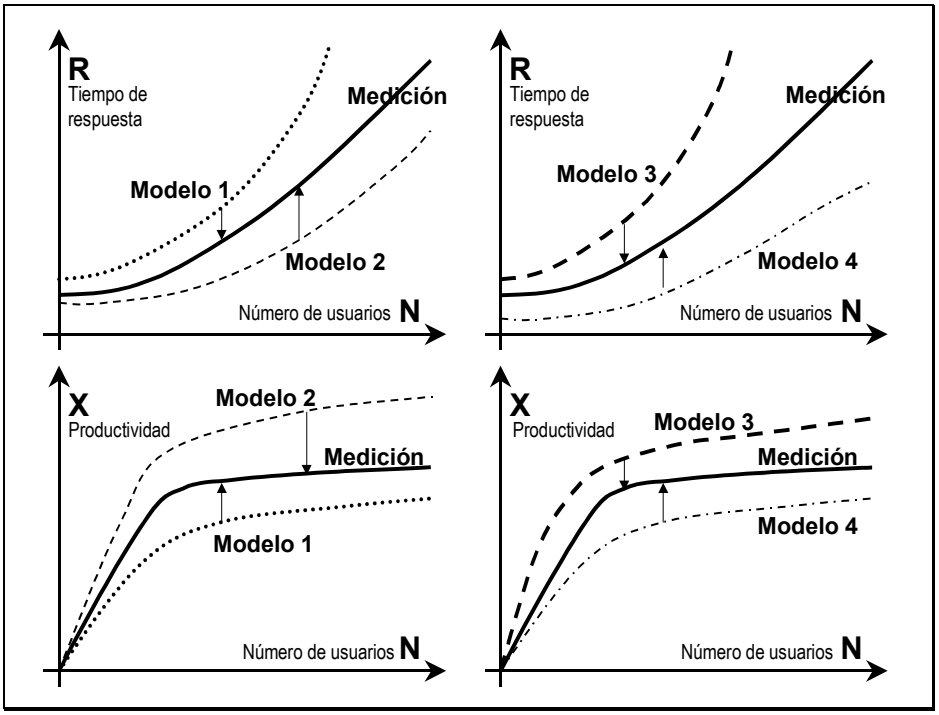
Por simplicidad, se van a fundir los procesos de validación y optimización, utilizando para ello todos los puntos donde se han realizado mediciones, y planteando el proceso como una concatenación de tres problemas simples del modo siguiente:

- Construye el modelo de redes de colas utilizando la herramienta *JSIMgraph* y especifica en este modelo todos los parámetros necesarios y que has calculado en los pasos previos. Especial atención a la pestaña de "*Routing section*" de las colas que representan la red y la CPU. Por defecto se asignan *Random*. Para que el modelo sea resoluble habrá que cambiarlo a *"Probabilities*" e introducir los valores de probabilidad de enrutamiento calculándolos a partir de las razones de visita obtenidas en el punto anterior. En la estación CPU, en la pestaña *Queue Section*, en el desplegable correspondiente a *Station queue policy* elige la opción *Procesor Sharing*, o procesador compartido, que es como trabajan realmente las CPU.
- Se resuelve el modelo invocando a la herramienta *JMVA* para todos los puntos (número de usuarios) en los que se realizaron mediciones (se realizará la resolución individual o con ayuda de la pestaña "*What-if*" según los casos). Se representan gráficamente los valores obtenidos con el modelo y se comparan con los valores obtenidos de la medición.
- Representar y comparar las curvas de utilizaciones de los componentes para el modelo y los valores reales. 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 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 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 anterior, 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. Aportar una tabla con los valores de los tiempos de servicio iniciales y finales, si se han modificado.**

# 4. Presentación de resultados

En la memoria de la práctica debe describirse, de forma completa y detallada, el proceso de construcción y ajuste de los dos modelos descritos.

- 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 según lo indicado en el paso 6 del apartado 2. En todos los casos indicar los tiempos de servicio utilizados (Tiempo de servicio de partida, tiempo de servicio obtenido del análisis y, si procede, tiempo de servicio final tras el ajuste). Conclusiones sobre la calidad del ajuste del modelo, si es mejorable o no, y si lo fuera qué propondrías.
- 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 optimización y validación del mismo según lo indicado en el paso 3 del apartado 3. Deben aparecer en una tabla los valores de partida a partir de los cuales se realizan los cálculos (demandas y razones de visita) así como los resultados obtenidos (tiempos de servicio y probabilidades de transición). Conclusiones sobre la calidad del ajuste del modelo, si es mejorable o no, y si lo fuera qué propondrías. Si se ha realizado un ajuste de los tiempos de servicio, mostar en una nueva tabla los nuevos valores utilizados.

ADEMÁS de la memoria, se entregarán en el campus virtual, para cada caso: los archivos JMT realizados, así como los archivos Excel que permitar seguir el proceso de ajuste y optimización.

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