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? 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: 5, 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 podría 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?

Para llevar a cabo este análisis sería necesario realizar nuevos experimentos modificando los parámetros que caracterizan las peticiones. Otra alternativa será construir un modelo de comportamiento del servidor y realizar las pruebas sobre el modelo, que será el objetivo de sucesivas prácticas.

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 *servidor_SCES*, 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 *servidor_SCES* consume CPU, lee y escribe información de/en disco y usa memoria. Los valores medios que caracterizan el consumo de recursos de cada transacción se le indican a la aplicación *servidor_SCES* en la línea de comandos cuando se la arranca. También se tendrá en cuenta el tiempo de reflexión, Z, utilizado por el inyector.

Cada equipo de trabajo realizará la prueba en condiciones distintas. Existirán unos parámetros fijos para todos los casos y otros específicos para cada equipo según se indican en las tablas siguientes:

TABLA 1.- Equipos situados en el lado de la mesa del profesor (2 procesadores)									
CPU	**Lect**	**Escr**	**Mem**	**Z(seg)**	**PL1**	**PL2**	**PL3**	**PL4**	**PL5**
85000	40	40	400	1,1				1	
85000	40	40	400	2,0			1		
85000	40	240	400	1,1					7
85000	40	240	400	2,0	1				
85000	240	40	400	1,1			4		
85000	240	40	400	2,0				**7**	
85000	240	240	400	1,1				5	
85000	240	240	400	2,0	7				
340000	40	40	400	1,1					3
340000	40	40	400	2,0				3	
340000	40	240	400	1,1			2		
340000	40	240	400	2,0	3				
340000	240	40	400	1,1		**1**			
340000	240	40	400	2,0			3		
340000	240	240	400	1,1		3			
340000	240	240	400	2,0					5

TABLA 2.- Equipos situados en el lado de la puerta del aula (4 procesadores)									
CPU	**Lect**	**Escr**	**Mem**	**Z(seg)**	**PL1**	**PL2**	**PL3**	**PL4**	**PL5**
110000	90	90	400	1,1					2
110000	90	90	400	2,4			8		
110000	90	360	400	1,1				2	
110000	90	360	400	2,4	4				
110000	360	90	400	1,1			7		
110000	360	90	400	2,4		8			
110000	360	360	400	1,1					6
110000	360	360	400	2,4		6			
440000	90	90	400	1,1	2				
440000	90	90	400	2,4			6		
440000	90	360	400	1,1				6	
440000	90	360	400	2,4		2			
440000	360	90	400	1,1	6				
440000	360	90	400	2,4					4
440000	360	360	400	1,1			5		
440000	360	360	400	2,4				4	

En la tabla las columnas etiquetadas como **CPU**, **Lect**, **Escr** y **Mem**, se refieren a los valores de los cuatro primeros parámetros con los que cada equipo lanzará el servidor. El quinto parámetro del servidor ,**Nº Ite**, será fijo para todos los equipos y tendrá un valor de **6** El número de usuarios dependerá de las condiciones de cada equipo. La carpeta de trabajo, tratará de ser siempre `C:\trabajo`. Por tanto, cada equipo invocará el servidor de la siguiente forma:

```
servidor_SCES CPU Lect Escr Mem 6 XXX C:\Trabajo
```

Donde los valores CPU, Lect, Escr y Mem se sustituirán por los correspondientes a cada fila de la tabla.

Cada equipo procederá de la siguiente forma:

- Dependiendo de su ubicación en el aula, elegirá una de las dos tablas anteriores. Los alumnos que realicen las prácticas en los equipos situados en el lado del profesor consultarán la Tabla 1, mientras que los que realicen las prácticas en los equipos situados en el lado de la puerta del aula, elegirán la Tabla 2.
- Posteriormente, en la tabla correspondiente, el equipo buscará en la columna etiquetada con su identificador de grupo de prácticas (PLX), la fila en la que se encuentre su identificador de

equipo de trabajo dentro del grupo. Así por ejemplo el equipo 1 del grupo de prácticas 7, situado en el aula en el lado de la mesa del profesor, consultará en la Tabla 1 la columna PL5 y dentro de ella marcará la fila en la que aparezca el 7, en este caso la tercera fila. Los valores que aparecen a la izquierda son los que tendrá que considerar en el experimento de carga.

- En el ejemplo, la columna Z tiene como valor 1,1 seg, éste es el tiempo de reflexión promedio a usar en el inyector el equipo 7 del grupo PL5.
- Los valores CPU=85000, Lect=40, Escr=240 y Mem=400 serán los primeros cuatro parámetros con los que se arrancará el servidor el equipo 7 del grupo PL5.

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. Por tanto, en la máquina donde se vaya a ejecutar el servidor *servidor_SCES* se creará el conjunto de recopiladores de datos necesario.

Por último, en la máquina en la que se ejecutará el programa *inyector*, éste se invocará con el tiempo medio de reflexión que se le haya asignado en las tablas anteriores. El intervalo de calentamiento, será un minuto, el intervalo de medición del experimento será de un **mínimo de 5 minutos** y el número de usuarios ira variando para realizar las pruebas de carga.

RECUERDA que no debe estar activa ninguna operación de E/S ni en inyector ni en el servidor mientras se está llevando a cabo la prueba de carga.

3. Presentación de resultados (Medición servidor)

Cada equipo de trabajo debe presentar:

- Una memoria en la que se desarrollen las tres tareas descritas a continuación, respondiendo a las preguntas planteadas, y aportando opiniones sobre el funcionamiento del sistema.
- Los archivos obtenidos de los experimentos realizados, así como el archivo resumen del análisis.

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. Utilizar todos los contadores de rendimiento citados en la práctica 2:

- Procesador: % tiempo de procesador.
- Memoria: Bytes disponibles, Bytes de cache, Pág./s, Errores de página/s.
- Disco Físico: % inactivo, Long promedio de la cola de disco, Transferencias de disco/s, Promedio en segundos/transferencia y Promedio de bytes por transferencia.
- Interfaz de red: Ancho de banda actual y Total de bytes/s.

A partir de la información recopilada de las pruebas se obtendrá:

- Representación gráfica de la evolución del tiempo de respuesta promedio (seg) y 90-percentil (seg).
- Representación gráfica de la evolución de la productividad promedio del servidor (peticiones/seg).
- Representación gráfica de la utilización media de los recursos del servidor: %CPU, %Memoria, %Disco y %Red.

Presentar esta información en forma de curvas utilizando como guía las figuras que aparecen posteriormente y responder a las preguntas que se indican relativas a los datos obtenidos.

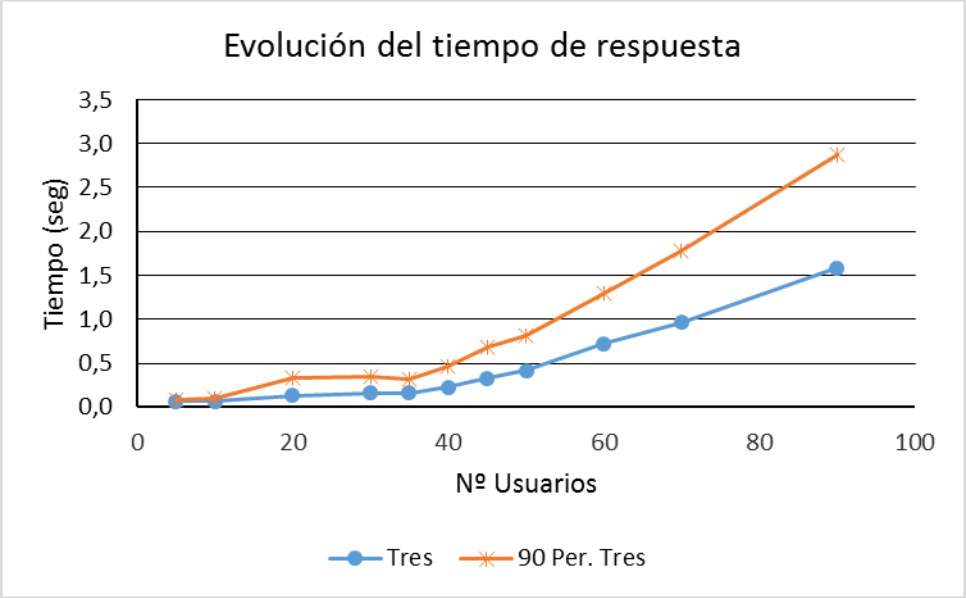
En función de las características de la transacción realizada por la aplicación *servidor_SCES*, para medir todos los regímenes de funcionamiento (productividad) del servidor (lineal, rodilla y saturación) el número máximo de usuarios a utilizar en los experimentos puede ser muy variable. Realizar experimentos para 1,2,3,4, … usuarios resulta excesivo. Para empezar, se recomienda realizar una experiencia con 5 usuarios. Esta experiencia permite determinar las características de la transacción realizada, esto es, el tiempo de respuesta de una transacción ejecutada con baja carga, 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. 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 de los parámetros asignados. Una vez localizados los regímenes se realizan experimentos adicionales para generar más puntos en las curvas. **En las gráficas deben aparecen resaltados los puntos donde se han realizado mediciones.**

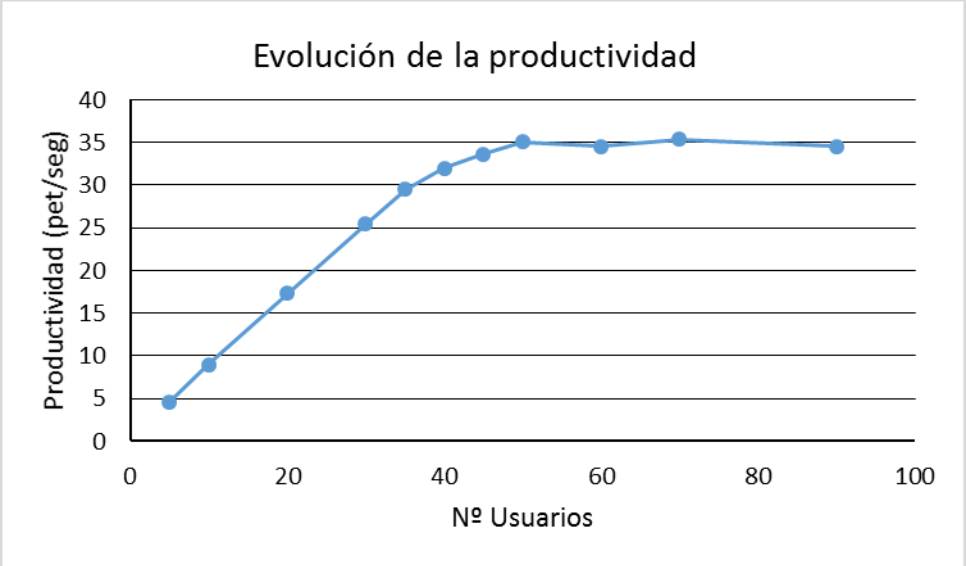
Además de las gráficas en la memoria, los resultados obtenidos de las mediciones deben incorporarse al archivo `PL`**`X`**`-EQ`**`Y`**`_CES_datos_practica3.xls`, donde X e Y deben sustituirse por el número de grupo de prácticas y equipo de trabajo respectivamente. Este archivo resumen debe formar parte del material a entregar con la práctica. También se incorporarán al material a entregar los archivos correspondientes a cada prueba a partir de los cuales se ha calculado el percentil y el uso de recursos.

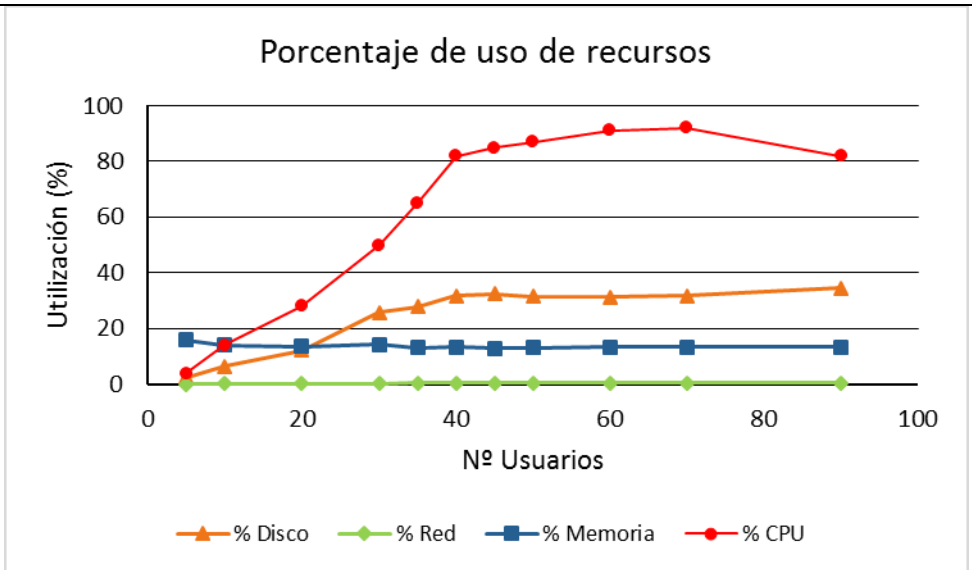
La no entrega de los ficheros supone una penalización de 3 puntos en la nota final de la práctica.

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.

Ejemplo de las gráficas a entregar







CUESTIONES RELATIVAS AL ANÁLISIS DEL FUNCIONAMIENTO DEL SERVIDOR

¿Cómo has calculado el % de uso de memoria?

Tomando como referencia la evolución de la productividad con el número de usuarios, ¿qué zonas de trabajo puedes diferenciar en el funcionamiento del servidor, 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 70% del recurso que primero se satura. Compara este valor con el obtenido para el punto de 5 usuarios.

Si se desea asegurar un tiempo medio de respuesta inferior al doble del tiempo para 5 usuarios, ¿cuántos usuarios simultáneos soportaría el servidor?

Si se desea asegurar que el 90% de las peticiones tengan un tiempo de respuesta inferior al doble del tiempo para 5 usuarios, ¿cuántos usuarios simultáneos soporta el servidor? Comenta la diferencia en el número de usuarios soportados, entre usar la media y el 90-percentil como métrica de la calidad del servicio.

¿Cuál es la máxima productividad absoluta que se puede obtener de este servidor y en que punto se alcanza? ¿Cuáles son los valores de tiempos de respuesta y las utilizaciones para ese punto? Compáralos con los valores correspondientes al punto de 5 usuarios, ¿son admisibles? ¿Por qué?

¿Cuál es el recurso que actúa como cuello de botella? ¿Cuál es su valor máximo de utilización? Si el sistema está en zona de saturación, y el valor de la utilización del dispositivo cuello de botella no alcanza niveles iguales o superiores al 90% como predice la teoría, ¿Cuál podría ser la causa?

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, al que denominaremos punto nominal, que estará situado al final de la zona lineal, pero sin entrar en la rodilla; y otro con 5 usuarios. **En estas pruebas no es necesario que mantengas activo el monitor de rendimiento de Windows.**

Para esos dos puntos, se llevará a cabo un análisis completo:

1. Para el punto nominal vamos a utilizar la técnica de las réplicas independientes. Realiza inicialmente 5 réplicas del experimento, **construye una tabla en la que debes anotar para cada una de estas réplicas su tiempo de respuesta y productividad promedio**. A partir de estos valores, calcula el número de réplicas que serían necesarias para que el tiempo de respuesta se pueda expresar con al menos una precisión del 10% con un nivel de confianza del 95%.
2. Para el punto con 5 usuarios se utilizará la técnica de la media por lotes. Realiza un experimento de medición largo, al menos equivalente a 4 veces la duración de un experimento normal. Guarda la información de las peticiones en un archivo Excel. Ordénalas por su tiempo de inicio. Inicialmente toma el tamaño del lote como la quinta parte del total de peticiones. Anota en una tabla el tiempo de respuesta y productividad promedio para cada lote. IMPORTANTE: cada lote tiene el mismo número de peticiones, pero no la misma duración temporal, EXPLICA cómo calcularías la productividad del lote. A partir de estos valores, calcular el número de lotes que serían necesarios para que el tiempo de respuesta se pueda expresar con al menos una precisión del 10% con un nivel de confianza del 95%. **Este archivo Excel formará parte del material a entregar.**

A partir de las medidas recopiladas en los puntos anteriores, calcular:

3. El intervalo de confianza obtenido para el tiempo de respuesta en cada punto (nominal y 5 usuarios).
4. Intervalo de confianza obtenido y errores cometidos para la productividad con una confianza del 95% sólo en el punto nominal. En este caso se parte de las 5 réplicas realizadas y se quiere calcular el error que se cometería. 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.
5. 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, en el punto nominal considerar las peticiones de todas las réplicas (agrupadas en una única hoja Excel) y en el punto de 5 usuarios tomar todas las peticiones del experimento largo. Indicar como se ha escogido el tamaño de celda para construir 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. **El archivo/os Excel utilizados para la realización del histograma también formará parte del material a entregar.**

TAREA 3: Considerando el histograma realizado para el **punto con 5 usuarios**, se llevará a cabo un proceso de ajuste de distribuciones. El objetivo será determinar qué distribución estadística sigue el tiempo de respuesta cuando existen pocos usuarios en el sistema. Este proceso constituye una aproximación para determinar el tiempo que realmente se emplea en atender una petición, cuando la petición dispone del sistema para ella sola y no ha de compartir recursos con ninguna otra.

1. A partir del histograma del tiempo de respuesta para el punto con pocos usuarios, y el análisis del resumen estadístico de sus datos (*herramientas → análisis de datos → estadística descriptiva* y el cálculo del coeficiente de variación), ¿podrías sugerir alguna familia de distribuciones como origen de los tiempos de respuesta medidos?. Toma como referencia las familias de distribuciones que aparecen en el documento "`Distribuciones_Continuas_Law&Kelton.pdf`" disponible en la carpeta `C:\Trabajo` de los equipos. **Nota:** Modificando la anchura de las clases de los histogramas el aspecto de este pueda orientar a una distribución u otra.

[OPCIONAL] Estudiar la representatividad del ajuste de distribución propuesto. La correcta realización de este apartado incrementará la calificación de la práctica en 2 puntos siempre y cuando no se supere la calificación máxima posible.