

Simulación y análisis del rendimiento de un servidor

Práctica 6

1. Objetivo

En esta práctica el alumno debe combinar los conocimientos y los datos obtenidos empleando las dos técnicas de análisis vistas (medición y modelado analítico), para, utilizando la técnica de evaluación de simulación, enriquecer la representatividad del modelo del sistema con objeto de conseguir un mayor ajuste de las predicciones a los resultados observados. Empleando la técnica de simulación, se estudiará la respuesta del sistema ante condiciones de trabajo distintas a las medidas y de esta forma se obtendrá una idea aproximada de cómo respondería el sistema ante la carga propuesta.

En esta práctica se considerarán extensiones mediante simulación en los dos tipos de modelos de sistema desarrollados:

(1) En el modelo a nivel de sistema se introducirá una distribución estadística para el tiempo de servicio distinta de la exponencial, así como otras consideraciones orientadas a ajustar los resultados del modelo a los resultados medidos en todo el rango de medición.

(2) En el modelo a nivel de componente se analizará el impacto en el sistema de una carga adicional, procedente de un origen distinto.

En ambos casos, deberán considerarse los problemas de eliminación del transitorio y determinación de la duración adecuada de la simulación empleando las técnicas que se consideren adecuadas.

En futuras prácticas sobre configuración, 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. Modificaciones a nivel de sistema

A nivel de sistema, una actuación que puede ayudar a mejorar la representatividad del modelo desarrollado es el uso de una distribución de tiempo de servicio diferente a la exponencial. En el modelado analítico tradicional sólo los sistemas con un reducido número de distribuciones son resolubles, por ese motivo para probar otro tipo de distribución será necesario acudir a la simulación.

A partir del histograma de tiempos de respuesta obtenido para 5 usuarios, proponer una distribución para el tiempo de servicio del servidor. Para desarrollar este punto existen varias alternativas:

1. Construir una función en QNAP que implemente la función de probabilidad buscada (consultar anexo a esta práctica). Posteriormente en la parte correspondiente a la cláusula /SERVICE/ se evaluaría inicialmente el valor correspondiente a la función y posteriormente se consumiría el tiempo utilizando la función CST.

2. A partir de los datos obtenidos en el histograma de tiempos de respuesta para un usuario y con ayuda de la función HISTOGR, establecer una forma de evaluación del tiempo de servicio acorde con las frecuencias observadas. El consumo de tiempo se realizaría de nuevo utilizando la función CST. La función HISTOGR, se le pasan como parámetros dos listas de reales: los extremos de los intervalos, y la probabilidad de pertenecer a cada intervalo.

Valor:=HISTOGR((1.0, 2.0, 3.0, 4.0),(0.27, 0.33, 0.4));

3. Como alternativa más sencilla se podría utilizar una distribución exponencial aunque a partir de un valor mínimo (de entre todas las peticiones realizadas para el caso de 5 usuarios, el tiempo de respuesta mínimo medido). Para cada petición, se obtendría un valor correspondiente a una distribución exponencial con la media indicada, si el valor obtenido es inferior al mínimo, se sustituiría por el mínimo, posteriormente con ayuda de la función CST se consumiría el tiempo correspondiente.

Desarrollar el modelo a nivel de sistema con la nueva distribución de tiempo de servicio en el servidor. Evaluar el comportamiento del modelo del sistema, considerando el problema del transitorio y la duración de la simulación.

Para aquellos grupos cuya curva de productividad presente una tendencia descendente tras la rodilla, se podría tratar de ajustar este proceso de degeneración del sistema explicándolo a partir de una sobrecarga del sistema operativo.

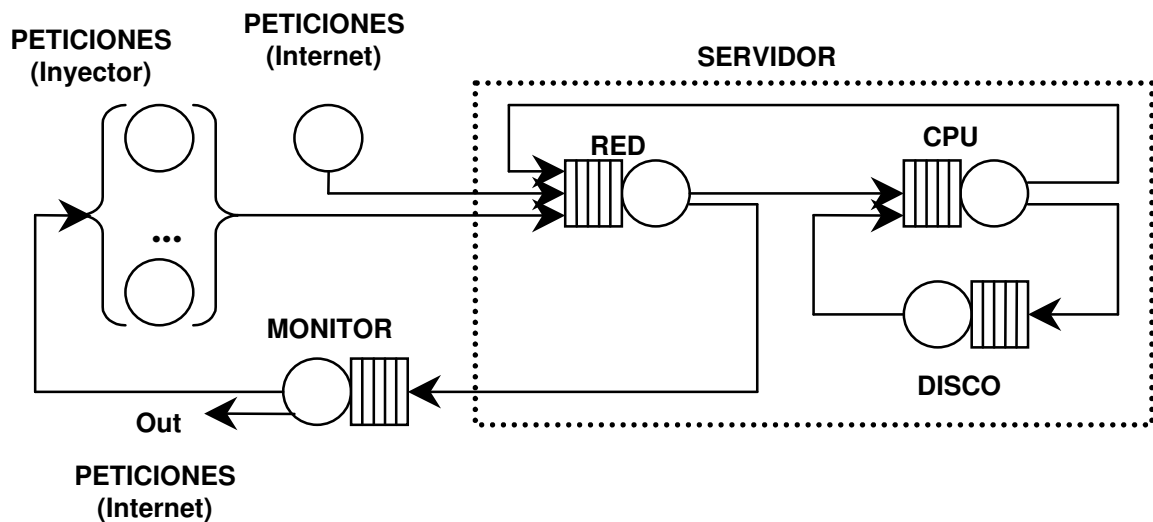
Se procedería:

1. Aplicando la ley de Little, obtener para cada punto medido el número de usuarios concurrentes (o peticiones simultáneas dentro del sistema). $N_c = X \times T_r$
2. En la curva de productividad, identificar el punto de máxima productividad (debería coincidir en el entorno de la rodilla) y anotarlo. Para ese punto se supone que el sistema está saturado, o lo que es lo mismo, su utilización es próxima al 100%. Por tanto aplicando la ley de Utilización: $U = X \times S$, como $U = 1$ el tiempo de servicio será la inversa de la productividad. Al tiempo de servicio en este punto llamadlo S_0 (su valor debe ser similar al que se habrá obtenido para el modelo a nivel de sistema en la práctica 6). A partir del punto de máxima productividad, obtener los valores de los tiempos de servicio utilizando la expresión: $S = (1/X) - S_0$.
3. Identificar a qué número de usuarios concurrentes se corresponde el valor de máxima productividad.
4. Con ayuda de la hoja Excel, construir un gráfico de dispersión de puntos, colocando en abscisas el número de usuarios concurrentes (**a partir del valor correspondiente a la máxima productividad**) y en ordenadas, los tiempos de servicio calculados para cada punto. En el gráfico de puntos, añadir una línea de tendencia e incluir la ecuación de la línea de tendencia. Anotar esta ecuación de tiempo de servicio en función del número de usuarios concurrentes. Se habrá obtenido de esta forma la sobrecarga del operativo cuando se supera un determinado número de usuarios o peticiones concurrentes.
5. Dentro de la estación servidor, evaluar el número de usuarios concurrentes (empleando la función CUSTNB) y si se supera el valor límite obtenido previamente, al tiempo de servicio normal habría que añadirle un tiempo de servicio adicional correspondiente a la sobrecarga del operativo. El valor de la sobrecarga vendría dado por la ecuación de la recta de regresión obtenida para el número de usuarios concurrentes.

3. Modificaciones a nivel de componentes

Una vez desarrollado el modelo de simulación para el sistema se va a proceder a realizar el análisis del comportamiento del sistema bajo unas condiciones de carga diferentes a las medidas.

Supondremos que nuestro sistema, aparte de recibir peticiones del inyector, recibe también peticiones enviadas desde otros computadores a través de Internet. Las peticiones son análogas a las enviadas por el inyector, con lo cual recibirán el mismo procesamiento en el servidor. El modelo conceptual de colas a representar será:



Construir el modelo de simulación para este sistema, incorporar las modificaciones necesarias para obtener las métricas de prestaciones a nivel global (productividad y tiempo de respuesta).

Si se supone que las peticiones de internet llegan según una distribución exponencial a razón de 2 peticiones/seg, ¿cuál sería la respuesta del sistema?

¿Cuál sería la cadencia de peticiones máxima, procedentes de Internet, que soportaría el sistema de forma que alcanzara la saturación con la mitad del número de usuarios del punto nominal?

4. Presentación de resultados

1. Para el estudio a nivel de sistema:

- Indicar cuál de las tres alternativas de representación del tiempo de servicio has utilizado y los pasos seguidos para implementarla y reajustar el modelo. Mostrar el código QNAP.
- Para el punto de funcionamiento de la práctica 3, realiza la simulación y muestra la evolución de las métricas de prestaciones de forma gráfica. A partir de los resultados determina las acciones a tomar para reducir el transitorio y establecer la duración de la simulación que produzca los resultados adecuados. ¿Cuál es la duración del transitorio? ¿Cuál será la duración de la simulación o el número de réplicas necesarias?
- Para el punto de 5 usuarios, obtén el tiempo de respuesta de todas las peticiones realizadas por el sistema. Construye el histograma de tiempos de respuesta y compáralo, en la misma gráfica, con el que habías obtenido en la práctica de medición.
- Simular el funcionamiento del sistema para los puntos medidos en el sistema real y obtener las métricas de prestaciones (productividad y tiempo de respuesta). Para cada métrica, representar sobre una misma gráfica los valores correspondientes a: los valores medidos en el sistema real, los resultados obtenidos del modelo analítico a nivel de sistema y los valores obtenidos con el modelo de simulación. Comentarios sobre los resultados presentados.
- Simular el funcionamiento del sistema para los puntos medidos en el sistema real y obtener las métricas de prestaciones (productividad y tiempo de respuesta). Para cada métrica, representar sobre una misma gráfica los valores correspondientes a: los valores medidos en el sistema real, los resultados obtenidos del modelo analítico a nivel de sistema y los valores obtenidos con el modelo de simulación. Comentarios sobre los resultados presentados.

2. Para la parte de componentes:

- Código QNAP para el modelo de simulación que implementa las peticiones a través de Internet. Incluir la parte correspondiente a la eliminación del transitorio y duración adecuada de la simulación.
- Simular el funcionamiento del sistema para los puntos medidos en el sistema real y obtener las métricas de prestaciones (productividad, tiempo de respuesta y utilización de dispositivos). Para cada métrica, representar sobre una misma gráfica los valores correspondientes a los valores medidos en el sistema real y los nuevos valores obtenidos con el modelo de simulación. ¿Cuál es el impacto de la nueva carga en el sistema?
- Valor de la cadencia de llegada de peticiones de Internet que cumple la condición pedida.

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

Anexo: Generación de funciones de distribución

Para obtener valores que sigan otras distribuciones estadísticas diferentes a la uniforme se puede utilizar el método de la transformación inversa. Este método permite generar variables aleatorias a partir de los valores obtenidos en el intervalo $[0, 1]$ de forma uniforme. Para ello supone que se quiere generar una variable aleatoria X que es continua y que tiene una función de distribución F que es continua y estrictamente creciente en el intervalo $(0, 1)$, esto es, si se cumple que $x_1 < x_2$ y que $0 < F(x_1) \leq F(x_2) < 1$ entonces $F(x_1) < F(x_2)$. Si se denota por F^{-1} a la inversa de la función F , entonces un algoritmo para generar la variable aleatoria X con función de distribución F es el siguiente:

1. Generar U que siga una distribución uniforme en $(0,1)$.
2. Retornar $X = F^{-1}(U)$.

Se sabe que $F^{-1}(U)$ siempre estará definida porque $0 \leq U \leq 1$ y el rango de F es $[0, 1]$.

Así, por ejemplo, para generar una variable X que siga una distribución exponencial de media β , se haría del siguiente modo. La función de distribución de una exponencial es la siguiente:

$$F(X) = \begin{cases} 1 - e^{-X/\beta} & \text{si } X \geq 0 \\ 0 & \text{en otro caso} \end{cases}$$

por lo tanto para encontrar F^{-1} , hacemos $U = F(X)$ y despejamos X , con lo que se obtiene lo siguiente:

$$X = F^{-1}(U) = -\beta * \ln(1 - U)$$

Por lo tanto, para generar la variable aleatoria deseada X se genera primero U , distribuida uniformemente en $(0,1)$ y entonces se hace $X = -\beta * \ln(U)$.

Se ha sustituido $(1-U)$ por U porque ambas variables siguen la misma distribución $U(0,1)$. Sin embargo, hay que tener en cuenta que esto provoca que haya una correlación negativa de las X s respecto a las U s.

En la siguiente tabla se muestran las transformaciones que hay que realizar sobre una distribución uniforme en el intervalo $(0,1)$, esto es, $U(0,1)$ que se denotará por U , para poder obtener diferentes distribuciones estadísticas siguiendo el método de la transformación inversa:

Distribución Estadística	Modo de obtenerla
Uniforme en $[A, B]$	$X = A + U * [B - A]$
Exponencial (β)	$X = -\beta * \ln(U)$
Weibull (α, β)	$X = \beta * [-\ln(U)]^{(1/\alpha)}$
Pareto (α)	$X = 1/U^{1/\alpha}$

Para obtener valores que sigan una **distribución normal de media μ y desviación típica σ** , esto es, $N(\mu, \sigma^2)$, se puede utilizar la función que proporciona QNAP a tal efecto, o se puede calcular a partir de la $U(0,1)$, utilizando transformaciones más complejas que para el resto de distribuciones. En primer lugar se debe trasladar la distribución uniforme a una distribución normal

de media 0 y desviación típica 1, $N(0, 1)$, y, posteriormente, se realizaría la siguiente operación para llevar ese valor a una $N(\mu, \sigma^2)$:

$$N(\mu, \sigma^2) = \mu + \sigma * N(0, 1)$$

Se debe tener en cuenta que la distribución normal lleva como parámetro la varianza, σ^2 , que es el cuadrado de la desviación típica σ .

Para obtener el valor de la distribución $N(0,1)$ se deberían seguir los siguientes pasos:

En primer lugar, se deberían obtener dos valores aleatorios distribuidos uniformemente en el intervalo (0,1), que se denotarán como U_1 y U_2 . Es importante tener en cuenta que se deben utilizar semillas diferentes para obtener los dos valores, ya que sino el segundo valor sería dependiente del primero, y entonces este método no sería válido.

A partir de los dos valores U_1 y U_2 se pueden obtener dos valores de la distribución $N(0,1)$ aplicando las siguientes operaciones:

$$\text{Valor_1} = \sqrt{-2 * \ln(U_1)} * \cos(2 * \pi * U_2)$$

$$\text{Valor_2} = \sqrt{-2 * \ln(U_1)} * \sin(2 * \pi * U_2)$$

Para obtener valores que sigan una **distribución lognormal** de media μ y desviación típica σ , $LN(\mu, \sigma^2)$, se deben obtener previamente valores que sigan una distribución normal de media μ_1 y desviación típica σ_1 , $N(\mu_1, \sigma_1^2)$, donde μ_1 y σ_1 se calculan a partir de μ y σ del siguiente modo:

$$\mu_1 = \ln(\mu^2 / \sqrt{\sigma^2 + \mu^2})$$

$$\sigma_1^2 = \ln[(\sigma^2 + \mu^2) / \mu^2]$$

Una vez obtenidos los valores de la $N(\mu_1, \sigma_1^2)$, que denotamos por Y , se les aplican las siguientes transformaciones para obtener los valores que pertenecen a la distribución $LN(\mu, \sigma^2)$:

$$LN(\mu, \sigma^2) = e^Y$$

Se pueden obtener valores que sigan muchas más distribuciones estadísticas realizando transformaciones de este tipo sobre los valores de la distribución $U(0, 1)$.