Examen Enero 2022

Ingeniería de Servidores

Ismael Sallami Moreno

Universidad de Granada
Doble Grado en Ingeniería Informática y ADE
https://elblogdeismael.github.io

2.- (1 punto)

Un servidor de base de datos en equilibrio de flujo recibe una media de 120 consultas por minuto. Sabemos que su disco duro tarda, de media, 30ms en atender cada petición de E/S que le llega (48ms si incluimos la espera en la cola) y que su productividad es 25 peticiones de E/S completadas por segundo. Calcule:

a) El número medio de peticiones de E/S en la cola de espera del disco duro (0,5 puntos).

b) ¿Cuánto tiempo, de media, consumen los accesos al disco duro por cada consulta que se realiza al servidor? (0,5 puntos)

a) Ya que el servidor esta en equilibrio de flujo, podemos usar la ley de Little, además lo que me esta pidiendo es el *Número medio de trabajos en la cola de espera.*

$$Q_{dd} = \lambda_{dd} \times W_{dd}$$
$$Q_{dd} = X_{dd} \times (R_{dd} - S_{dd}) =$$
$$25\frac{tr}{s} \times \frac{48-30}{1000}s = 0,45tr = 0,45 \text{ peticiones de E/S}$$

b) Nos esta preguntando por la D_{dd}.

$$U_{dd} = X_{dd} \times S_{dd} = 25\frac{tr}{s} \times 30\frac{ms}{tr} = 0,75$$
$$D_{dd} = \frac{B_{dd}}{C_{dd}} = \frac{B_{dd}/T}{C_{dd}/T} = \frac{U_{dd}}{X_{dd}} = \frac{0,75}{120\frac{tr}{min}} = 0,00625min = 0,375s$$

3.- (2 puntos)

Considere un servidor web que es modelado con los siguientes parámetros:

Dispositivo	S_i (ms)	V_i
CPU	10	16
Disco A	20	7
Disco B	30	8

Conteste de forma razonada a estas preguntas **indicando claramente las definiciones y/o leyes operacionales que ha necesitado utilizar**:

a) Calcule la productividad máxima que puede llegar a alcanzar la CPU en el seno de este servidor (0,5 puntos).

b) Estime la utilización del disco A si el servidor web recibe una media de 5 peticiones por segundo (0,5 puntos).

c) Suponiendo que el servidor se encuentre en equilibrio de flujo, encuentre una expresión que permita calcular el tiempo medio de respuesta de la CPU en función de su utilización y su tiempo de servicio (pero no hace falta que calcule el valor numérico) (0,5 puntos).

d) Si el servidor web recibe una media de 2 peticiones por segundo, calcule el número medio de clientes conectados al servidor, suponiendo que cada cliente envía un único trabajo al servidor (0,5 puntos).

a) Los pasos a seguir son:

$$V_i = \frac{C_i}{C_0} = \frac{C_i/T}{C_0/T} = \frac{X_i}{X_0} \Rightarrow X_i = V_i \times X_0 \text{ (ley de flujo forzado)}$$

$$\text{Por lo que para max nos queda: } X_{max} = V_{max} \times X_0$$

$$\text{Sabemos que estamos en saturación: } X_0^{max} = \frac{1}{D_b}$$

Sabiendo esto, debemos de calcular las demandas en cada caso, aplicando la fórmula de $D_i = V_i \times S_i$,y nos queda que son Dcpu = 0,16s, DdiscoA es 0,14s y DdiscoB es de 0,24s. Entonces el cuello de botella es el discoB, por ende, nos queda:

$$X_0^{max} = \frac{1}{0,24} = 4,16667 tr/s$$

Lo que buscamos es:

$$X_{cpu}^{max} = X_0^{max} \times V_{cpu} = 4,17 \times 16 = 66,72 \text{ procesos ejecutados por segundo}$$

b) En el apartado anterior hemos visto que el servidor se satura cuando la tasa media de llegada supera los $4,2$ tr/s. Por tanto, el servidor estará saturado con 5 tr/s. Así que la utilización del disco A nunca podrá superar su valor máximo, que se puede calcular de la siguiente forma:

- Relación utilización - demanda de servicio: $U_i = X_0 \times D_i$
- Por tanto, $U_{i,max} = X_{0,max} \times D_i$

En particular, para el disco A:

$$U_{discoA,max} = X_{0,max} \times D_{discoA} = 4,17 \text{ tr/s} \times 0,14 \text{ s} = 0,58 \ (58\,\%)$$

Y esa es la estimación que podemos hacer de cuál será la utilización del disco A con esa tasa de llegada.

c) La fórmula que nos piden es la siguiente:

$$R_i = \frac{S_i}{1 - U_i}$$

d) Para una tasa de llegada de 2 tr/s $< X_0^{max}$, sabemos que el servidor está en equilibrio de flujo, por lo que $X_0 = \lambda_0 = 2$ tr/s y podemos utilizar la ley de Little aplicada al servidor completo:

$$N_0 = X_0 \times R_0$$

donde N_0 es el número medio de trabajos en el servidor (número medio de clientes conectados al servidor suponiendo que cada cliente envía un único trabajo al servidor).

Por tanto, solo nos queda calcular R_0 para poder resolver el ejercicio. Como estamos en equilibrio de flujo, podemos utilizar la ley general del tiempo de respuesta:

$$R_0 = V_{cpu} \cdot R_{cpu} + V_{discoA} \cdot R_{discoA} + V_{discoB} \cdot R_{discoB}$$

Aprovechando la expresión que hemos obtenido en el apartado anterior:

$$R_i = \frac{S_i}{1 - U_i} = \frac{S_i}{1 - X_0 \cdot D_i}$$

Por tanto:

$$R_{cpu} = 0,015 \ s \quad R_{discoA} = 0,028 \ s \quad R_{discoB} = 0,058 \ s$$

Obtenemos $R_0 = 0,89 \ s$.

Y finalmente:

$$N_0 = 2 \ tr/s \times 0,89 \ s = 1,78$$

Es decir, hay 1,78 clientes conectados al servidor de media.

4.- (1.25 puntos)

En *Google* están intentando mejorar la técnica de distribución de carga de sus servidores de *YouTube*. Para ello, han realizado 100 medidas de la productividad media de los servidores durante un número determinado, pero fijo, de horas para las 2 configuraciones principales de distribución de carga: *Conf1* y *Conf2*. Como los experimentos se han realizado en presencia de aleatoriedad, han realizado un test-t cuyos resultados son:

Paired Samples T-Test

Measure 1	Measure 2	t	df	p	Mean diff.	90 % CI for Mean Diff.	
Conf1	*Conf2*	0.113	99	0.91	0.88	Lower: -19.5	Upper: 21.3

A partir de esta información, conteste **de forma razonada** a las siguientes cuestiones, indicando explícitamente qué valores concretos de la tabla anterior son los que necesita y cómo los usa:

a) Si las diferencias fuesen significativas, ¿qué configuración presentaría el mejor rendimiento, utilizando como criterio la media aritmética? (0,5 puntos)

b) ¿Cuál es la hipótesis de partida de este test-t? ¿Hay diferencias significativas en el rendimiento para un 99 % de nivel de confianza? (0,5 puntos)

c) En general, explique cuál es la diferencia entre $\overline{d}$ y $\overline{d}_{\text{real}}$ cuando se realiza un test t para evaluar si los rendimientos entre dos alternativas son estadísticamente diferentes. (0,25 puntos)

a) Vemos que el valor medio de la diferencia entre las productividades de Conf1 menos las de Conf2 es un número positivo (Mean difference = 0,88). Por tanto, la productividad media conseguida por Conf1 es mayor que la de Conf2. Como tiene mejor rendimiento el que consigue la mayor productividad, sería Conf1 la mejor configuración.

b) La hipótesis nula (H0) del test-t es que el rendimiento medio de Conf1 es igual al de Conf2, es decir, no hay diferencias significativas entre ambas configuraciones. El nivel de confianza es del 99 %, lo que implica un valor de $\alpha = 0{,}01$. El p-valor obtenido es 0,91, que es mucho mayor que α. Por tanto, no podemos rechazar la hipótesis nula y concluimos que no existen diferencias significativas en el rendimiento para un 99 % de nivel de confianza.

c) La primera variable se refiere a la media muestral de la diferencia entre los rendimientos, es decir, la que se calcula de las muestras experimentales que hemos obtenido, mientras que la segunda es la media real de dicha diferencia, es decir, el verdadero valor medio de la diferencia entre los rendimientos si no hubiese aleatoriedad en las medidas o el valor al que debería tender la media muestral si hiciéramos infinitas medidas.

5.- (1 punto) Cuestiones.

a) ¿Para qué se usa una "rack unit" (1U) en el mundo de los servidores? (0,25 puntos)

b) ¿Qué diferencia hay entre los conceptos de precisión y exactitud cuando hablamos de la medida realizada por un sensor? (0,25 puntos)

c) ¿Cómo obtiene `gprof` información sobre el número de veces que se ha ejecutado cada función de un programa? ¿Es un valor estimado o exacto? (0,25 puntos)

d) Indique las principales características de `perf` (0,25 puntos).

a) Una *rack unit* (1U) es una unidad estándar de medida utilizada para definir la **altura** de los dispositivos que se montan en racks de servidores. 1U equivale a **1,75 pulgadas** o **44,45 mm** de alto. Permite organizar de forma compacta múltiples servidores y equipos en un mismo bastidor, optimizando el espacio y la refrigeración.

b) **Exactitud** (*accuracy*) indica cuánto se desvía el valor medido respecto al valor real (error sistemático). **Precisión** (*precision*) indica la consistencia entre varias mediciones del mismo valor real. Es posible que una medición sea precisa pero no exacta, y viceversa.

c) `gprof` obtiene el número de veces que se ha ejecutado cada función mediante instrumentación del código fuente, insertando contadores en cada función. Este número de llamadas es un **valor exacto**. El tiempo de ejecución de cada función, sin embargo, es un valor estimado basado en muestreo estadístico del contador de programa.

d) `perf` es una herramienta avanzada de análisis de rendimiento en Linux. Permite monitorizar eventos del sistema (CPU, cachés, interrupciones, llamadas al sistema), analizar el rendimiento de procesos o del sistema completo, utilizar contadores de hardware y eventos del kernel, y generar perfiles detallados para detectar cuellos de botella y puntos calientes en aplicaciones.

6.- (0.75 puntos)

Tras vender un riñón para cambiar nuestra vieja tarjeta gráfica por una flamante `Nvidia RTX 3090 Founders Edition`, ahora un programa de un solo hilo que usa solamente CPU y GPU tarda 3 veces menos que antes. Pero además, la parte del programa que hace uso de la nueva tarjeta gráfica ahora (= *tras la mejora*) tarda un tercio del tiempo de ejecución actual (= un tercio del tiempo de ejecución mejorado). ¿Qué fracción del tiempo de ejecución original era la usada por la tarjeta gráfica antigua?

Para resolver el ejercicio, llamemos T_o al tiempo de ejecución original y T_m al tiempo tras la mejora. Nos piden la fracción f de T_o que el hilo usaba la GPU antes de la mejora.

Sabemos que:

$$T_m = \frac{T_o}{3}$$

y que la parte del programa que usa la GPU tras la mejora tarda un tercio del tiempo de ejecución mejorado:

$$T_{\text{gpu,m}} = \frac{T_m}{3} = \frac{T_o}{9}$$

Sea $T_{\text{gpu,o}}$ el tiempo que la GPU ocupaba antes de la mejora y T_{cpu} el tiempo de CPU (que no cambia). Entonces:

$$T_o = T_{\text{cpu}} + T_{\text{gpu,o}}$$

$$T_m = T_{\text{cpu}} + T_{\text{gpu,m}}$$

Despejando T_{cpu} de la primera ecuación:

$$T_{\text{cpu}} = T_o - T_{\text{gpu,o}}$$

Sustituimos en la segunda:

$$T_m = (T_o - T_{\text{gpu,o}}) + T_{\text{gpu,m}}$$

$$\frac{T_o}{3} = T_o - T_{\text{gpu,o}} + \frac{T_o}{9}$$

Despejamos $T_{\text{gpu,o}}$:

$$T_o - T_{\text{gpu,o}} + \frac{T_o}{9} = \frac{T_o}{3}$$

$$T_o + \frac{T_o}{9} - \frac{T_o}{3} = T_{\text{gpu,o}}$$

$$T_o \left(1 + \frac{1}{9} - \frac{1}{3}\right) = T_{\text{gpu,o}}$$

$$T_o \left(\frac{9}{9} + \frac{1}{9} - \frac{3}{9}\right) = T_{\text{gpu,o}}$$

$$T_o \left(\frac{7}{9}\right) = T_{\text{gpu,o}}$$

Por tanto, la fracción pedida es:

$$f = \frac{T_{\text{gpu,o}}}{T_o} = \frac{7}{9} \approx 0{,}78$$

Respuesta: La fracción del tiempo de ejecución original que usaba la GPU antes de la mejora era $\boxed{\frac{7}{9}}$ (aproximadamente 0,78).