Examen Enero 2023

Ingeniería de Servidores

Ismael Sallami Moreno

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

2.- (1 punto) Cuestiones (0,25 puntos cada pregunta).

a) ¿Para qué sirve la pasta/cola térmica cuando estamos montando un servidor?

La pasta/cola térmica se utiliza para mejorar la transferencia de calor entre el procesador y el disipador, asegurando que el procesador se mantenga a una temperatura adecuada durante su funcionamiento.

b) ¿Qué son los módulos de memoria del tipo LR-DIMM y qué ventaja aportan?

LR-DIMM significa Load Reduced DIMM. Hay un buffer que almacena tanto las señales de control como los datos a leer/escribir. Tienen mayor latencia que los R-DIMM, pero permiten módulos de mayor capacidad. Además, incorporan corrección de errores (ECC).

c) ¿De qué palabras proceden las siglas SSD?

Las siglas SSD proceden de "Solid State Drive", que se traduce como Ünidad de Estado Sólido". Son dispositivos de almacenamiento que utilizan memoria flash para guardar datos, ofreciendo mayor velocidad y resistencia a golpes en comparación con los discos duros tradicionales (HDD).

d) ¿Qué significa .ºLTPçuando hablamos de benchmarks como TPC-C? ¿Qué hacen ese tipo de benchmarks?

OLTP significa "Online Transaction Processing", que se traduce como "Procesamiento de Transacciones en Línea". Los benchmarks como TPC-C miden el rendimiento de sistemas de bases de datos en entornos de transacciones en línea, evaluando su capacidad para manejar múltiples transacciones simultáneas y su eficiencia en la gestión de datos.

3.- (0,5 puntos)

Un computador tarda 300 segundos en ejecutar un programa. El 66 % del tiempo se utiliza en operaciones en el procesador, mientras que el resto se dedica a acceder a dispositivos de Entrada/Salida. ¿Cuántas veces tendrá que mejorar el procesador si queremos que el programa se ejecute 6 veces más rápido? ¿Cuál es la ganancia máxima que podría obtenerse mejorando solo el procesador?

Sea $T_o = 300$ segundos el tiempo original de ejecución del programa.

La fracción del tiempo que se dedica al procesador es $f = 0{,}66$, y el resto $(1-f = 0{,}34)$ se dedica a dispositivos de E/S.

Queremos que el programa se ejecute 6 veces más rápido, es decir, que el nuevo tiempo T_m cumpla:

$$S = \frac{T_o}{T_m} = 6$$

La fórmula de la Ley de Amdahl es:

$$S = \frac{1}{(1-f) + \frac{f}{k}}$$

donde k es el factor de mejora del procesador. Sustituimos los valores conocidos:

$$6 = \frac{1}{0{,}34 + \frac{0{,}66}{k}} \Rightarrow 0{,}34 + \frac{0{,}66}{k} = \frac{1}{6} \Rightarrow \frac{0{,}66}{k} = \frac{1}{6} - 0{,}34 = \frac{1 - 2{,}04}{6} = -\frac{1{,}04}{6}$$

Esto da un valor negativo, lo cual indica que no es posible alcanzar una ganancia de 6 solamente mejorando el procesador.

Ganancia máxima posible:
La ganancia máxima se obtiene cuando $k \to \infty$ (es decir, el procesador se mejora infinitamente). En ese caso, la parte del procesador se anula y:

$$S_{\text{máx}} = \frac{1}{1-f} = \frac{1}{0{,}34} \approx 2{,}94$$

Para que el programa se ejecute 6 veces más rápido es necesario mejorar también la parte de Entrada/Salida. La ganancia máxima que puede obtenerse mejorando únicamente el procesador es aproximadamente **2,94**.

4.- (0,5 puntos)

Se sabe que el monitor *sar* de un determinado servidor consume, cada vez que se activa, 150ms de tiempo de CPU y 210KiB de DRAM. Sabiendo que nuestro equipo solo tiene una CPU y 2GiB de DRAM, calcule cada cuánto tiempo, en segundos, debe activarse el monitor para que la sobrecarga de la CPU sea 0,8 % **(0,25 puntos)**. Calcule igualmente

la sobrecarga (en tanto por ciento) de DRAM del monitor cada vez que éste se activa **(0,25 puntos)**.

Datos:

- Tiempo de CPU por activación: $T_{\text{cpu}} = 150\,\text{ms} = 0{,}15\,\text{s}$
- Memoria usada por activación: $M = 210\,\text{KiB}$
- Memoria total del sistema: $M_{\text{total}} = 2\,\text{GiB} = 2 \times 1024 \times 1024 = 2\,097\,152\,\text{KiB}$
- Sobrecarga deseada de CPU: $U = 0{,}8\,\% = 0{,}008$

a) Intervalo entre activaciones para una sobrecarga de CPU del 0,8 %

La sobrecarga se define como:

$$\text{Sobrecarga CPU} = \frac{T_{\text{cpu}}}{T_{\text{activación}}} \Rightarrow T_{\text{activación}} = \frac{T_{\text{cpu}}}{\text{Sobrecarga}} = \frac{0{,}15}{0{,}008} = 18{,}75\,\text{segundos}$$

El monitor debe activarse cada **18,75 segundos** para que la sobrecarga de CPU sea del 0,8 %.

b) Sobrecarga de DRAM por activación

$$\text{Sobrecarga DRAM} = \frac{210\,\text{KiB}}{2\,097\,152\,\text{KiB}} \times 100 \approx 0{,}01\,\%$$

La sobrecarga de DRAM es de aproximadamente **0,01 %** cada vez que se activa el monitor.

5.- (0,75 puntos)

En *Samsung* están intentando evaluar la mejora en la latencia de los módulos de memoria DRAM que introduce la nueva tecnología DDR5. Para ello, han realizado 100 experimentos para calcular las latencias medias en múltiples diferentes contextos. Finalmente, para comprobar que las diferencias en las latencias entre un módulo DDR4 y otro DDR5 no se deben a efectos aleatorios, han realizado un test t, cuyos resultados son los que aparecen en la siguiente tabla (las latencias han sido medidas en *ns*):

Measure 1	Measure 2	t	df	p	Mean difference
DDR4	DDR5	0.13	99	0.91	0.88

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 tecnología presentaría el mejor rendimiento, utilizando como criterio la media aritmética? **(0,25 puntos)**

Si las diferencias fueran significativas, **la tecnología con mejor rendimiento sería la que presenta menor latencia media**.

Según la tabla, la **diferencia media** es de **0,88 ns** (última columna), lo que significa que:

$$\text{Media DDR4} - \text{Media DDR5} = 0{,}88 \Rightarrow \text{DDR5 tiene menor latencia media.}$$

Por tanto, si la diferencia fuera significativa, **DDR5 presentaría un mejor rendimiento** que DDR4 al ofrecer menor latencia.

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)**

La hipótesis nula (H_0) de un test-t para diferencias de medias es:

$$H_0 : \mu_{\text{DDR4}} = \mu_{\text{DDR5}}.$$

Es decir, no hay diferencias significativas entre las latencias medias de DDR4 y DDR5.

El resultado del test t muestra un **valor p = 0,91** (columna "p"), que es muy superior al umbral típico $\alpha = 0{,}01$ para un **nivel de confianza del 99 %**.

Dado que:

$$p = 0{,}91 \gg 0{,}01,$$

no podemos rechazar la hipótesis nula.

No hay evidencia suficiente para afirmar que las diferencias en latencia entre DDR4 y DDR5 son estadísticamente significativas con un 99 % de nivel de confianza.

6.- (2,5 puntos)

El informático responsable de una empresa dedicada a juegos de azar ha modelado el servidor web que atiende a los clientes utilizando técnicas de análisis operacional. Este modelo está formado por una CPU y dos discos. Los valores medios de los parámetros relevantes del mismo se muestran a continuación:

Dispositivo	**Tiempo de servicio (ms)**	**Razón de visita**
CPU (1)	10	15
Disco A (2)	35	6
Disco B (3)	40	3

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

a) Calcule el tiempo de respuesta medio mínimo de este servidor **(0,25 puntos)**.

Ley utilizada: Tiempo de respuesta mínimo:

$$R_{\min} = \sum_{i=1}^{k} V_i \cdot S_i$$

$$R_{\min} = 15 \cdot 10 + 6 \cdot 35 + 3 \cdot 40 = 150 + 210 + 120 = \boxed{480\,\text{ms}}$$

b) Estime la utilización del Disco B si el servidor web recibe una media de 6 peticiones por segundo **(0,5 puntos)**.

si el servidor recibe una media de 6 tr/s, estará saturado ($6 > \text{x0}_{\max}$). por lo tanto, *no se cumple el equilibrio de flujo y no podemos identificar X_0 con la tasa de llegada.*

La utilización del disco B nunca podrá superar su valor máximo, que se puede calcular a partir de la relación utilización-demanda de servicio:

$$U_i = X_0^{\max} \cdot D_i$$

Por tanto, $\text{U}_{\max} = \text{X}_{0\max} \cdot \text{D}_\text{B}$. En particular, para el disco B:

$$U_{\text{discoB\_max}} = 4{,}76\,\text{tr/s} \cdot 0{,}12\,\text{s} = 0{,}57\,(57\,\%)$$

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

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

Una forma de resolverlo sería: partimos de la ley del flujo forzado:

$$X_i = X_0 \cdot V_i \quad \Rightarrow \quad X_{\text{imax}} = X_0^{\max} \cdot V_i$$

En particular, para el Disco A:

$$X_{\text{discoA\_max}} = X_0^{\max} \cdot V_{\text{discoA}} = 4{,}76 \cdot 6 = \boxed{28{,}6\,\text{tr/s}}$$

Otra forma: como el Disco A es el cuello de botella, puede llegar a alcanzar una utilización = 1. Usando la ley de la utilización:

$$U_i = X_i \cdot S_i \quad \Rightarrow \quad X_{\text{discoA}} = \frac{1}{S_{\text{discoA}}} = \frac{1}{0{,}035} = \boxed{28{,}6\,\text{tr/s}}$$

Algunos indican que por ser el Disco A cuello de botella:

$$X_{\text{discoA\_max}} = X_0^{\max} = \frac{1}{D_{\text{discoA}}}$$

Obviamente, eso es incorrecto (y un fallo grave).

d) Supuesto equilibrio de flujo, exprese el tiempo medio de respuesta de la CPU en función de su productividad y su tiempo de servicio **(0,5 puntos)**.

Se deben indicar todos los pasos necesarios para llegar a la siguiente expresión:

$$R_i = \frac{S_i}{1 - X_i \cdot S_i}$$

Y de ahí, particularizarlo para la CPU:

$$R_{\text{cpu}} = \frac{S_{\text{cpu}}}{1 - X_{\text{cpu}} \cdot S_{\text{cpu}}}$$

e) Calcule el número medio de procesos en la cola de la CPU si la tasa de llegada al sistema es de 4 req/s **(0,5 puntos)**.

Comprobamos primero que estamos en equilibrio de flujo: $4\,\text{tr/s} < X_{0\text{max}}$, por lo que $X_0 = \lambda_0 = 4\,\text{tr/s}$.

Aplicando la ley de Little a la cola de la estación de servicio de la CPU (válida porque estamos en equilibrio de flujo):

$$Q_{\text{cpu}} = X_{\text{cpu}} \cdot W_{\text{cpu}} = X_{\text{cpu}} \cdot (R_{\text{cpu}} - S_{\text{cpu}})$$

Donde R_{cpu} se puede obtener del apartado anterior. Calculamos primero X_{cpu}:

$$X_{\text{cpu}} = V_{\text{cpu}} \cdot X_0 = 15 \cdot 4 = 60\,\text{tr/s}$$

De aquí, $R_{\text{cpu}} = 0{,}025\,\text{s}$ (del apartado anterior) y $S_{\text{cpu}} = 0{,}01\,\text{s}$.

Por tanto,

$$Q_{\text{cpu}} = 60\,\text{tr/s} \cdot (0{,}025\,\text{s} - 0{,}01\,\text{s}) = 60 \cdot 0{,}015 = 0{,}9\,\text{procesos}$$

El número medio de procesos en la cola de la CPU es 0,9.

f) ¿Cuáles serían los nuevos valores de S_3 y V_3 si la CPU se reemplaza por otra el doble de rápida? **(0,25 puntos)**

Reemplazar la CPU por otra más rápida no influye ni en el tiempo de servicio ni en la razón de visita del Disco B. El tiempo de servicio (S_3) y la razón de visita (V_3) del Disco B dependen únicamente de las características del propio disco y del patrón de acceso de la aplicación, no de la velocidad de la CPU.

Por tanto:

$$S_3' = 40\,\text{ms}, \quad V_3' = \boxed{3}$$

7.- (0,75 puntos)

Queremos diseñar un servidor de ayuda a la docencia al que se conectarán unos 30 estudiantes durante las 2 horas que duran las sesiones de prácticas de la asignatura. Este servidor consta de una CPU, un disco duro y una tarjeta de red. Tras la prueba de funcionamiento de 2 horas con 30 estudiantes, se han medido los siguientes valores:

Dispositivo	**Tiempo de servicio (s)**	**Razón de visita**
CPU (1)	0,01	80
DISCO (2)	0,5	20
RED (3)	0,24	5

¿Cuánto tiempo debería transcurrir, de media, entre que un estudiante recibe la respuesta de este servidor hasta que vuelve a realizar una nueva petición, para que 30 sea precisamente el número ideal de clientes de este servidor?

Ley utilizada: En redes de colas cerradas con N clientes y tiempo de reflexión Z, el número de usuarios conectados cumple:

$$N = X \cdot (R + Z)$$

donde:

- $N = 30$ es el número ideal de clientes,
- R es el tiempo medio de respuesta del servidor,
- Z es el tiempo medio entre que el cliente recibe la respuesta y vuelve a hacer una petición (tiempo de reflexión),
- X es la productividad del sistema (número de peticiones atendidas por segundo).

Para que N sea ideal, el sistema debe estar en régimen óptimo, por lo que usamos la productividad máxima:

$$X_{\max} = \frac{1}{D_{\max}} \quad \text{donde } D_i = V_i \cdot S_i$$

Calculamos las demandas de servicio:

$$D_1 = 80 \cdot 0{,}01 = 0{,}8 \text{ s} \quad D_2 = 20 \cdot 0{,}5 = 10 \text{ s} \quad D_3 = 5 \cdot 0{,}24 = 1{,}2 \text{ s}$$

El cuello de botella es el **disco**, por tanto:

$$X_{\max} = \frac{1}{10} = 0{,}1 \,\text{peticiones/s}$$

Calculamos el tiempo de respuesta total:

$$R = \sum V_i \cdot S_i = 0{,}8 + 10 + 1{,}2 = 12\,\text{s}$$

Despejamos Z de la fórmula:

$$30 = 0{,}1 \cdot (12 + Z) \Rightarrow 300 = 12 + Z \Rightarrow Z = \boxed{288\,\text{s}}$$

Para que el número ideal de clientes sea 30, el tiempo de reflexión medio debe ser de **288 segundos**.

Nota: para saber cuando es de cola cerrada o abierta, debemos de tener en cuenta:

- Cerrada: número de clientes fijo, no cambia, en este caso 30 estudiantes.
- Abierta: número de clientes variable, puede cambiar, como en un servidor web donde los usuarios entran y salen continuamente.