

## Problema 2 Temas 2, 4, 5, 6.

Se está planteando un sistema de control de tráfico en una zona de acceso restringido. El sistema de almacenamiento necesario tendrá un volumen muy elevado de accesos sobre el fichero de 'vehículos' (aprox, unos  $5 \cdot 10^5$  registros de 168 bytes reales cada uno). Las claves de búsqueda (selección) en este registro son  $K_1$ ,  $K_2$ , y  $K_3$ , de 12, 17, 15 bytes de media cada una, respectivamente. Se tendrá un soporte direccionado cuyo tamaño de bloque es de 2KB. Se precisa tomar una decisión acerca de la organización más adecuada con el objetivo de optimizar el rendimiento en número de accesos a soporte.

En esta decisión se quiere tener en cuenta el conjunto de procesos  $P = \{P_1, P_2, P_3\}$ , cuyas frecuencias relativas son  $f(P) = \{60\%, 20\%, 20\%\}$ , y su descripción:

- $P_1$ : Modificación de registros: selección por clave de identificación  $K_1$ ; las claves  $K_1$  y  $K_2$  no alteran su contenido (pero la clave  $K_3$  si).
- $P_2$ : Consulta por clave no unívoca  $K_2$  (de media 25 registros por valor).
- $P_3$ : Selección por clave no unívoca  $K_3$  (625 valores distintos).

El rendimiento del resto de los procesos no es crítico, y por ello no serán tenidos en cuenta para la decisión final, aunque deberá señalarse como inconveniente la degeneración de la organización, en su caso. Se dispone de una transformación sobre  $N=65.000$  para  $K_1$  que arroja una tasa de desbordamientos del 0,1%. Las alternativas que se manejan son todas organizaciones no consecutivas, con  $E_c=1$  y espacio libre distribuido para modificaciones (PCTFREE=10%). La tasa de modificación es  $T_M=30\%$  diaria, pero con el espacio libre distribuido, todos los registros modificados caben en el cubo donde estaban. Las propuestas son:

- $O_1$ : serial no consecutiva.
- $O_2$ : direccionada con  $CD = K_1$ , y gestión de desbordamientos serial.
- $O_3$ : secuencial no consecutiva con clave de ordenación física  $K_3$ .

### Se pide:

**a) Hallar el Coste Global inicial de cada organización (en número de accesos) y la densidad real de cada organización, sabiendo que la densidad ideal es del 92%.**

	<b><math>O_1</math>: serial nc.</b>	<b><math>O_2</math>: dir. <math>K_1</math></b>	<b><math>O_3</math>: sec. nc. <math>K_3</math></b>
<i>densidad real</i>	75,47%	58%	75,47%
<b><math>C(O_i, P_1)</math></b>	25000,5	1,03	25000,5
<b><math>C(O_i, P_2)</math></b>	50000	65.050 acc	50000
<b><math>C(O_i, P_3)</math></b>	50000	65.050 acc	97 acc
<b><math>C(O_i, P)</math></b>	35000,3 acc	26020,62 acc	25.019,7 acc

$$T_c = \lfloor 2048 \cdot 0,9 / 168 \rfloor = 10 \text{ reg/cubo}$$

$$T_{\text{util}} = d_i \cdot T_{\text{real}} = 0,92 \cdot 168 = 154,56 \text{ bytes}$$

$$O_1: N = 500.000 \text{ reg} / 10 \text{ reg/cubo} = 50.000 \text{ cubos}$$

$$d_r = 500.000 \cdot 154,56 / (50.000 \cdot 2048) = 75,47\%$$

$$O_2: \text{desb} = 500.000 \text{ reg} \cdot 0,1\% = 500 \text{ reg} \rightarrow 500 \text{ reg} / 10 \text{ reg/cubo} = 50 \text{ cubos desb}$$

$$d_r = 500.000 \cdot 154,56 / (65.050 \cdot 2048) = 58\%$$

$$n^\circ \text{ acc} = (499.500 \cdot 1 + 500 \cdot (1 + 51/2)) / 500.000 = 1,0255$$

$O_3$ : es una organización no consecutiva sin huecos  $\rightarrow$  misma densidad inicial que  $O_1$

$$500.000 \text{ rg} / 625 \text{ valores} = 800 \text{ rg/val}; + \text{un fallo: } 801 \text{ rg}/10 \text{ rg/cubo} = 81 \text{ cubos}$$

$$n^\circ \text{ acc} = \log_2 (50.000 + 1) + 81 = 97 \text{ acc.}$$

b) Calcular el **periodo de reorganización máximo en O<sub>3</sub>** para que su densidad no sea peor que la de O<sub>2</sub>. ¿Cuál es el coste de O<sub>3</sub> justo antes de la reorganización?

$$C'(O_3, P) = 35.519,7 \text{ acc}$$

¿Qué organización de las tres te parece más recomendable? Justifica tu respuesta.

$$T_M = 30\% \rightarrow 500.000 \cdot 30\% = 150.000 \text{ reg}$$

$$\rightarrow 150.000 \text{ rg} / 10 \text{ rg/cubo} = 15.000 \text{ cubos diarios}$$

$$d_r = 500.000 \cdot 154,56 / ((50.000 + 15.000 \text{ dias}) \cdot 2048) \geq 58\%$$

$$500.000 \cdot 15,456 / 58 \cdot 2048 \geq 50.000 + 15.000 \text{ dias}$$

$$65059,27 \geq 50.000 + 15.000 \text{ dias}$$

$$\text{días} \leq (65059,27 - 50.000) / 15.000 \rightarrow \text{días} \leq 1,004 \rightarrow \text{reorganización diaria (periodo 1 día)}$$

Antes de la reorganización,  $N = 50.000 + 15.000 = 65.000$  cubos

$$C'(O_3, P_1) = 65000 + 1 / 2 = 32500,5 \text{ acc}$$

$$C'(O_3, P_2) = 65000 \text{ acc}$$

$$C'(O_3, P_3) = 97 \text{ acc} + 15000 = 15097 \text{ acc}$$

$$C'(O_3, P) = 35.519,7 \text{ acc}$$

**Es recomendable, por tanto, la O<sub>2</sub>, porque la secuencial degenera mucho (al cabo de menos de tres horas ya sería peor que la direccionada).**

Suponiendo una degeneración lineal,  $(35519,7 - 25.019,7) / 24 = 437,5 \text{ +acc/hora}$

La diferencia  $(26020,62 - 25.019,7) / 437,5$  se supera en 2,288 horas (2 horas y 17,27 minutos).

c) Se opta por implantar la organización O<sub>2</sub> mejorada con la introducción de algunos índices para los procesos menos eficientes. Compara el rendimiento de un **índice serial** (no consecutivo) con el de otro en **árbol B<sup>+</sup>** ( $T_{\text{ptro\_int}} = 2B$ ) para cada clave involucrada.

**¿Qué índice/s se decide incluir? (sobre qué clave/s y qué tipo/s de índice)**

**Calcula el coste Global de O<sub>4</sub>** (O<sub>2</sub> más este/os índice/s)

Seriales:

**K<sub>2</sub>:**

$$K_2 = 17B + 1 + 25 \cdot 2 = 68B \rightarrow T_{c\_ind} = 2048 / 68 = 30 \text{ entradas/cubo}$$

$$500.000 \text{ reg} / 25 \text{ reg/ent} = 20.000 \text{ ent} \rightarrow 20.000 / 30 = 667 \text{ cubos seriales}$$

$$C(O_4, P_2) = 667 + 1 / 2 + 25 = 359 \text{ accesos}$$

**K<sub>3</sub>:**

$$500.000 \text{ rg} / 625 \text{ val} = 800 \text{ rg/val} \rightarrow K_3 = 15B + 2 + 800 \cdot 2 = 1617B \rightarrow T_{c\_ind} = 1 \text{ ent/cubo}$$

$$\rightarrow 625 \text{ cubos seriales}$$

$$C(O_4, P_3) = 625 + 1 / 2 + 800 = 1113 \text{ accesos}$$

Árboles B<sup>+</sup>:  $T_{\text{ptro\_int}} = 2B$

**K<sub>2</sub>:**

$$2m + (m-1) \cdot 17 \leq 2048 \rightarrow m \leq (2048 + 17) / 19 = 108 \rightarrow m_{\min} = 54$$

$$k \cdot 68 + 2 \leq 2048 \rightarrow k \leq 30,08 \rightarrow 30 \rightarrow k_{\min} = 15$$

$$\text{nivel } n: n^{\circ} \text{máx. hojas} = 20000 / 15 = 1333 \text{ hojas}$$

$$\text{nivel } n-1: 1333 / 54 = 24 \text{ nodos}$$

$$\text{nivel } n-2: 24 / 54 = 0,44 \rightarrow 1 \text{ nodo} \rightarrow \text{raíz} = 1 = n-2 \rightarrow n = 3 \text{ niveles}$$

$$C(O_4, P_2) = 3 + 25 = 28 \text{ accesos}$$

**K<sub>3</sub>:**

$$2m + (m-1) \cdot 15 \leq 2048 \rightarrow m \leq (2048 + 15) / 17 = 121 \rightarrow m_{\min} = 61$$

$$k \cdot 1617 + 2 \leq 2048 \rightarrow k \leq 1 \rightarrow k_{\min} = 1$$

nivel n:  $n^{\circ}\text{máx.hojas} = 625/1 = 625$  hojas  
nivel n-1:  $625/61 = 10$  nodos  
nivel n-2:  $10/61 = 0,16 \rightarrow 1$  nodo  $\rightarrow$  raíz = 1 = n-2  $\rightarrow$  n= 3 niveles  
 $C(O_4, P_2) = 3 + 800 = 803$  accesos

Por tanto, se eligen los índices en árbol B+ y el coste global queda:  
 $C(O_4, P) = 1,03*0,6 + 28*0,2 + 803*0,2 = 166,82$  accesos

d) Se observa que el proceso  $P_3$  tiene unas características peculiares: la condición se establece sobre  $K_3$ , y el proceso devuelve el valor de  $K_2$  para ese registro (el resto de campos no son relevantes). Se opta por implementar este proceso mediante un acceso invertido, para lo que se precisa contar con sendos índices invertidos sobre estas claves.

**¿Modificarías alguna de las decisiones tomadas en el apartado (c)?**

**¿Cómo queda la organización definitiva  $O_5$ ? (org. base + orgs. auxiliares)**

**¿Cuál es el coste global de la nueva organización  $O_5$ ?**

$K_2$ :

$2m + (m-1) \cdot 17 \leq 2048 \rightarrow m \leq (2048+17)/19 = 108 \rightarrow m_{\min} = 54$

$k \cdot (17+1+25 \cdot 4) + 2 \leq 2048 \rightarrow k \leq 17 \rightarrow 30 \rightarrow k_{\min} = 9$

nivel n:  $n^{\circ}\text{máx.hojas} = 20000/9 = 2222$  hojas

n-1:  $2222/54 = 41$  nodos

n-2:  $41/54 = 0,76 \rightarrow 1$  nodo  $\rightarrow$  raíz = 1 = n-2  $\rightarrow$  n= 3 niveles (igual)

$K_3$ : require hojas de 2 bq  $\rightarrow$  4 acc para recuperar la entrada

El acceso invertido ( $K_3 \rightarrow K_2$ ) requiere  $4 + 2222$  (hojas) = 2226 accesos más que el indizado, por ser muy grande el B+

Con un serial, habría sido

$K_2 = 17 B + 1 + 25 \cdot 4 = 118 B \rightarrow T_{c\_ind} = 2048/68 = 17$  entradas/cubo

$500.000 \text{ reg} / 25 \text{ reg/ent} = 20.000 \text{ ent} \rightarrow 20.000/30 = 667$  cubos seriales

$C(O_4, P_2) = 667 + 1 / 2 + 25 = 359$  accesos

### **Conclusiones:**

Con serial:  $10 + 800 = 810$  accesos

Con B =  $10 + 1667 = 1677$  accesos

En este proceso, se ganan 867 acc. con serial

En el  $P_2$ , se ganan 396.5 con árbol B

Como tienen la misma probabilidad, interesa que el índice sobre  $K_2$  sea **serial**

**$O_5$ : organización direccionada sobre  $K_1$ ,  $E_c=1$ ,  $N=65000$ ,**

**con índice serial no consecutivo sobre  $CI=K_2$  + índice en árbol B sobre  $C_I=K_3$**

**$C(O_5, P) = 1,03*0,6 + 425.5*0,2 + 810*0,2 = 0 + 400 = 88,42$  accesos**