Data Management – exam of 09/07/2020

Problem 1

If S is a schedule on transactions $T_1, \ldots, T_n$, then the *reduced precedence graph* $\rho(S)$ associated to S is a graph that has the transactions in S as nodes, and has an edge from T_i to T_j if and only if S contains two conflicting actions $a_i(x)$ in T_i and $a_j(x)$ in T_j on the same element x such that $a_i(x)$ precedes (not necessarily directly) $a_j(x)$ in S, and there is no action $\alpha_k(x)$ in S between $a_i(x)$ and $a_j(x)$, with k different from i and j. Prove or disprove the following claims:

1. If $\rho(S)$ is cyclic, then S is not conflict serializable.
2. If $\rho(S)$ is acyclic, then S is conflict serializable.
3. If $\rho(S)$ is cyclic, then S is not view serializable.

Problem 2 Consider the following schedule S:

$$r_1(x)\ r_2(x)\ w_3(x)\ w_3(z)\ c_3\ r_4(z)\ w_4(y)\ c_4\ w_1(y)\ c_1\ r_2(y)\ c_2$$

1. Is S a 2PL schedule with both shared and exclusive locks? Motivate your answers in detail.
2. Describe the behavior of the timestamp-based scheduler when processing S, assuming that, initially, for each element α of the database, we have rts(α)=wts(α)=wts-c(α)=0, and cb(α)=`true`, and assuming that the subscript of each action denotes the timestamp of the transaction executing such action.

Problem 3 Let `R(A,B,C,D)` be a relation stored in a heap file with 89.100 pages, and consider the following query:

```
select A, count(*) from R group by A
```

We have two options for executing the query: (1) We execute it at a single processor P having 300 buffer frames available. (2) We carry out a parallel execution of the query using n processors (with $n > 0$), each one with 16 buffer frames available. You are asked to tell if there is a value for n such that strategy (2) is more efficient (with respect to the elapsed time) than strategy (1). If the answer is positive, then tell which is the minimum value n such that the above condition holds, and tell which is the corresponding cost, in terms of both the total number of page accesses, and the elapsed time. If the answer is negative, then explain why. In both cases, motivate your answer in detail.

Problem 4

Consider the relation `City(code,name,region,country,population,mayor)`, with 1.000.000 tuples, the relation `Theater(city,tcode,size)`, that contains, for each value of the attribute `city`, an average of 10 tuples with that value, and the following query Q:

```
select name, region, avg(size)
from City, Theater
where code = city
group by region, name
order by region, name
```

We know that 300 values fit in one page of our system, and that our buffer has 50 free frames. Tell which physical structures (including possible indexes) you would choose for storing the two relations in such a way to optimize the above query Q, and tell the cost of executing the query under the assumptions that the two relations are stored according the chosen method, and all values occupy the same space. Motivate your answers in detail.

Problem 5

Let $\mathtt{R_1(A,B,C,D,E,F)}$ be stored in a heap file with 840.000 tuples, and $\mathtt{R_2(G,H,L)}$ be stored in a heap file, with 9.000.000 tuples and with an associated B$^+$-tree index on attribute `G`. We assume that every value has the same size, that every page has room for 600 values, that $V(\mathtt{R_1,B}) = 100$, $V(\mathtt{R_1,F}) = 300$, and that we have 125 free buffer frames available. Consider the following query:

```
select distinct B, G, L
from R1, R2
where F = G and L > 10
```

and answer the following questions:

1. Illustrate the logical plan associated to the above query expression.
2. Describe the selected logical plan, motivating the annswer.
3. Describe the physical plan you would choose, and determine the cost of executing the query according the chosen physical plan, motivating the answer.