Data Management – exam of 12/07/2021

Problem 1

A schedule S is called "one-writer" if at most one transaction appearing in S has write actions. Prove or disprove the following two statements:

1. A "one-writer" schedule S is conflict-serializable if and only if it is view-serializable.
2. A "one-writer" schedule S is conflict-serializable if and only if it is a 2PL schedule.

The solution must be provided under the assumption that no transaction reads or writes the same element more than once.

Problem 2

Consider the following schedule S:

$$B(T_0)\ w_0(A)\ c_0\ B(T_1)\ w_1(A)\ B(T_2)\ r_2(A)\ w_2(B)\ c_1\ B(T_3)\ r_3(A)\ w_3(C)\ w_3(B)\ w_2(C)\ c_3\ c_2$$

where the action B means "begin transaction", and every write action performed by transaction T_i writes the value i on the element corresponding to the argument. Suppose that S is executed by PostgreSQL with all the transactions defined with the isolation level "read committed", and for each action different from B tell what is the effect of the action and what is the behavior of the system when it executes the action.

Problem 3

Let `R(A,B)` be a relation with 100.000 pages, and `S(C,D)` a relation with 500.000 pages. We know that (*i*) `R` has 2.000 values in `A`, uniformly distributed in the various tuples, (*ii*) there is a clustered, sparse B$^+$-tree index on `R` with search key `A`, (*iii*) 60 tuples of `R` fit in one page, (*iv*) we have 260 frames available in the buffer, and (*v*) every value and pointer occupy the same amount of space. Consider the query:

```
select A,D
from R, S
where B = C and A >= 10 and A <= 15
```

Show the logical query plan associated to the query, as well as the logical query plan and the physical query plan you would choose for executing the query efficiently. Also, tell which is the cost (in terms of number of page accesses) of executing the query according to the chosen physical query plan.

Problem 4

Let `Building(bcode,floors,area,value,cat,city)` be stored in a heap file with 840.000 tuples, and `City(ccode,nation,nab)` be stored in a heap file with 9.000.000 tuples. We assume that every value has the same size, that every page has room for 600 values, that $V($`Building,floors`$) = 100$, $V($`Building,city`$) = 300$, and that we have 85 free buffer frames available. Consider the query:

```
select distinct floors, city, nab
from Building, City
where city = ccode and nab > 10000
```

Show the logical query plan associated to the query, as well as the logical query plan and the physical query plan you would choose for executing the query efficiently. Also, tell which is the cost (in terms of number of page accesses) of executing the query according to the chosen physical query plan.

Problem 5

If R_1`(A,B)` and R_2`(A,B)` are two relations each with key `A`, then the *disjoint left-union* of R_1 and R_2, indicated as $R_1 \oslash R_2$, is the relation with attributes `A,B` and with the set of tuples specified as follows: (*i*) for each tuple $t \in R_1$ such that t.B is not null, we have $t \in R_1 \oslash R_2$; (*ii*) for each tuple $t \in R_1$ such that t.B is null and such that there exists $t' \in R_2$ with $t.A = t'.A$, we have $t' \in R_1 \oslash R_2$. Suppose that we have a multiprocessor system with N nodes $n_1, \ldots, n_N$, each of them with $M > N$ free frames available, and that the two relations R_1 and R_2 are stored in node n_1.

1. Illustrate a parallel algorithm for computing $R_1 \oslash R_2$.
2. Assuming $B(R_1) = 10.000$, $B(R_2) = 15.000$, $N = 10$ and $M = 40$, describe the cost of the algorithm both in terms of the elapsed time, and in terms of number of page accesses.