CS 5/7330 Fall 2022

Final Exam

Answer all questions. There is a total of 100 points, full mark is 89. Please upload your answer as a PDF file to Canvas before the deadline (12/12 3:00pm).

1. (20 points) Consider the following schedule of 5 transactions:

Time	T1	T2	T3	T4	T5
1		x2 = Read(X)			
2					x5 = Read(X)
3				y4 = Read(Y)	
4				y4 = y4 * 2	
5	z1 = Read(Z)				
6		x2 = x2 + 1			
7			z2 = Read(Z)		
8					x5 = x5 * 2
9	z1 = z1 * 3				
10		y2 = Read(Y)			
11			z3 = z3 - 1		
12				Write(Y, y4)	
13		Write(X, x2)			
14	Write(Z, z1)				
15		y2 = y2 + x2			
16				x4 = Read(x)	
17					Write(X, x5)
18				x4 = y4 + x4	
19			Write(Z, z3)		
20		Write(Y, y2)			
21				Write(X, x4)	

Use the same convention as of homework 2. With the initial value of X, Y, Z in the database = 1, 2, 3 respectively. Trace the execution of this set of transaction, with the following convention (similar to homework 2):

i. Every time a transaction wants to read/write an item in the disk for the first time, it will request a share/exclusive lock (respectively). We denote it as the operation S-Lock(), X-Lock() respectively
ii. If a request is granted, the transaction can proceed immediately and execute the corresponding command
iii. Otherwise, the transaction will use the wait-die policy to determine whether it will wait or abort.
iv. *Notice that for the purpose in (iii), we denote the "time" of the transaction by the transaction number (e.g,. T1 has time 1 etc.).*

v. If a transaction waits, it will wait until the lock is released and attempt to obtain it. It will try to obtain the lock immediately after it was released by another transaction. If it can obtain the lock, it will execute all commands up to the time that the lock is requested/obtained.
vi. If a transaction aborts, it will not be restarted
vii. If there is more than one transaction waiting for the same lock, the lock is always granted to the transaction T_i with the smallest value of i.
viii. We assume a transaction immediately commits and release all the locks after the last operation is successfully executed.

Update the table (using the same convention of homework 2), to show the execution of the schedule. Also show the final values of X, Y, Z, as well as the corresponding serial schedule.

2. (18 points)_Consider we want to apply linear hashing to the following list of numbers:

   16, 22, 3, 7, 1, 15, 2, 19, 9, 4, 21, 0, 8,

   Assume the following:
   - We use the same hashing function as discuss in class (last k digits of the binary expression of the number – where k depend son the current level of the hash table).
   - Each bucket can hold 2 numbers
   - We split when the next bucket to be split when the bucket we insert the number to overflows (or is already overflowing)
   - Overflow buckets are used if a bucket overflows but it is not its turn to split
   - We start at level 0 (with 1 bucket)

   Show the content of the hash table ***after every split***. For each instance, you should show the following

   - Content of each bucket (show overflow buckets as linked list).Label the bucket number in binary.
   - Which bucket is to be split next
   - Example:

     Bucket 000 : (a, b), (c, d), (e, f), (g, h), (i)
     Bucket 001 : empty
     Bucket 10 : (j, k)
     Bucket 11 : (l)
     Bucket 100: (m, n), (o, p)
     …..
     Next bucket to be split : ??
     (a-p are actual data in the buckets, you need to show the value).
     (How numbers are ordered within a bucket does not matter).

3. (14 points) Consider we have three tables A, B, C, stored on a magnetic hard drive. The following are the information about the tables:

- Each page is 2000 bytes.
- No tuples can be split between pages.
- Size of each table:

	A	B	C
# of tuples	20,000	30,000	60,000
Bytes per tuple	100	200	300

- Table A is stored on one track of the disk
- Table B is evenly split between two tracks (different from the tracks that store A or C)
- Table C is evenly split between two tracks (different from the tracks that store A or B)
- Data on each track is stored as contagious pages
- Assume each seek + rotation takes 50ms, and each page read/write takes 0.5ms
- For every query in this question, we assume the disk head is at neither of the tracks that store a table in the beginning.

a. What is the number of pages needed to store each table? (You can assume there is no storage overhead)

b. Now consider the following query (assume we do NOT write the result back to the disk):

   *SELECT **
   FROM C
   WHERE C.d = 100

   Suppose we have a non-clustering index on C.d (which is NOT part of any key attributes). Suppose k tuples are being returned.

   i. If we do sequential scan. Calculate the time for this query.
   ii. Now suppose we use the non-clustering index, and we are required to retrieve the tuples in the order of how one access the index. Calculate the time for this query in the worst case.
   iii. Find the range of k such that (i) is a faster way to execute the query then (ii)

   (You may need to write your answers in part (i) and (ii) in terms of k).

4. (22 points) (Using the same setup as question 3. For this question, ignore seek time/rotational delay in your calculations).
   Consider a query that joins 3 tables, A, B and C with the join condition A.x= B.x and B.x = C.x.
   Assume the following:
   - A.x, B.x, C.x are all integer attributes
   - None of the attributes is the primary key of each table
   - Every value that appears in B.x appears in A.x (but not necessarily vice versa)
   - A.x has values between 1- 1000 (inclusive) evenly distributed.
   - C.x has values between 201 – 2200 (inclusive), evenly distributed,

- No extra information about distribution of B.x is available.

Consider the following

a. Suppose we want to join table A and C first. What should be the join condition?
b. Assume we use nested loop join to join table A and C. Which table should be in the outer loop? Why?
c. Now assume we have 200 buffers available. Further assume we evenly distribute the buffers between the two tables. Calculate the time needed to execute this query. Show your work.
d. How many tuples will the result have?
e. To make calculation simple, assume the join result of A and C contains every attribute (including duplicating the join attributes). Now assume we store the result of the join onto the disk. How many pages are needed?
f. Now suppose we have a outpuf buffer of 200 pages (i.e. the join store the results in the buffers, writing happens only when the output buffer if full, and all pages will be written to the same track. There will be a didcated track (different from tracks that store the tables for the result). Calculate the time needed to write the output.
g. Now suppose we join B with the result from part (e) that is written to the disk, using nested loop. Once again, we split the 200 buffers evenly between the two tables. Calculate the time for this join.
h. Calculate the overall time for the query. Remember to factor in the time you write the result of the first join to the disk. (But you can assume you are not writing the final result to the disk).

5. (18 points) Consider you want to use histograms to store value of an integer attribute D.x. Assume the following is the values of D.x (there is total of 50 tuples) (Assume you know that the range of the attribute is from 1 to 30)

1	2	2	2	2	2	2	2	2	2
4	7	7	7	8	8	11	11	11	12
13	14	15	16	16	16	16	16	17	17
17	17	17	18	20	24	25	25	25	25
26	26	27	27	28	28	29	29	30	30

a. Consider we want to build an equal-width histogram with 10 columns. Show the histogram using a table form (as described below:)

Range	Frequency
1-3	9
4-6	2
<Fill in the rest of the rows>	

b. Now consider we want to use a equi-depth histogram with 10 columns (so frequency of each column is 5). Show the histogram using the same format as (a)

c. Nos suppose you ONLY have the histogram available to estimate the number of tuples of the following query:

```
SELECT *
FROM D
Where d.x >= p and d.x <= q
```

Use each histogram to estimate the number of tuples returned for each of the following values of p,q. (round the result to the nearest integer if necessary)

i. p = 2, q = 2

ii. p = 16, q = 30

iii. p = 11, q = 16

6. (8 points) What conditions of ACID will one violate if it does not follow the write-ahead log principle? Illustrate with example(s).