**Department of Mathematics**
**I.I.T., Kharagpur**

**End Autumn Semester Examination 2016-2017**
**Subject :File Organization & Database Systems(MA40004/MA61018/MA60050)**
**4th Yr. B.Tech., 4th Yr. M.Sc., 1st Yr. M.Tech(CSDP)**
**No. of Students (115 + 29)**

**Time : 03 Hours** **Marks : 50**

Answer all **FIVE** Questions
(This question paper consists of **THREE** Pages)

1.

a) Draw a $B^+$ tree for a file of records with key values shown below:

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5 | 8 | 12 | 16 | - | 20 | 36 | - | 51 | 56 | - | 70 | 78 | 84 | 90 | 97 | - | 120 | 148 | 180 |

Assume that each block can contain at most three records and internal nodes can contain two key values and corresponding addresses. Now perform the following operations in sequence:

i) insert a record with key value 75.
ii) Delete a record with key value 51.

Also write one algorithm to search a record with a given key value from the existing $B^+$ tree.

b) Discuss the advantages and disadvantages of transferring data bucket wise instead of block wise for sequential and random processing of data. For a fixed set of operations on a file, how optimal bucket size can be determined?

c) Define i) average seek time , ii) average rotational latency time , iii) block transfer time and iv) effective block transfer time.

(4M+3M+3M)

2.

a) Consider a hash index which uses the hash function $h(x) = x$, and in which each hash bucket can hold 2 data items. The hash index is initially empty and contains 4 buckets (i.e., it uses the last two bits of $h(x)$ initially, although no data items are currently in the hash index). Now **draw the hash table** for the following sequence of insertions of records (insert only key value) into the hash index: 15, 63, 0, 16, 31, 47, 32, and 48. If the above hash index uses **extendible hashing**, show the final state of the index after the insertions.

b) Consider the following database scheme:
Client (client_no, client_name, address); Project (project_no, project_name, client_no);
Consultant (consultant_no, consultant_name) and Assignment (consultant_no, project_no,rate);
Express the following queries in **SQL** and **QUEL** :

i) Find the consultant names who are associated either in Project "P1" or in Project "P2".
ii) Find the client numbers and the corresponding rate of the project who have not chosen Mr. XYZ as consultant.
iii) Find the name of the project whose rate is highest.

(4M+6M)

3.

a) Consider the following schedule for the transactions T1, T2 and T3:

| T1 | T2 | T3 |
|---|---|---|
| | Read(Z) | |
| | Read(Y) | |
| | Write(Y) | |
| | | Read(Y) |
| | | Read(Z) |
| Read(X) | | |
| Write(X) | | |
| | | Write(Y) |
| | | Write(Z) |
| | Read(X) | |
| Read(Y) | | |
| Write(Y) | | |
| | Write(X) | |

i) Draw the precedence graph. Whether the schedule is serializable? Justify your answer.

ii) Design a schedule of T1, T2 and T3 which is serializable.

iii) Let *TS(T)* is the time stamp value of the transaction T. If *TS(T1)< TS(T2)<TS(T3)*, can the schedule given in the problem be executed under time stamp ordering protocol? If no, write a new schedule following time stamp ordering protocol.

b) Consider the relation BOOK (title, author, pname, lc_no), BORROWER (name, addr, city, card_no) and LOAN (card_no, lc_no, date). Draw the initial operator graph for the query

$$\pi_{title}(\sigma_{date<'24/4/2012'}(XLOANS))$$

where

$$XLOANS = \pi_S(\sigma_F(LOAN \times BORROWER \times BOOK)),$$

*F= BORROWER.Card_no=LOAN. Card_no ∧ BOOK. lc_no=LOAN. lc_no and S=title, author, pname, lc_no, name, addr, city, card_no, date.*

Apply heuristic rules to transform the query into a more efficient optimized form. Show all the intermediate steps.

(2M+2M+3M+3M)

4.

a) Consider the following transaction database for the set of items {A,B,C,D,E,F,H}:

| Trans ID | Items Purchased |
|---|---|
| T100 | A,B,C,D |
| T200 | A,B,C,E |
| T300 | A,B,E,F,H |
| T400 | A,C,H |

Suppose that minimum support is set to 50% and minimum confidence to 60%. Apply Apriori algorithm to find all frequent item sets together with their support. For all frequent itemsets of maximal length, list all corresponding association rules satisfying the requirements on minimum confidence together with their confidence.

b) Consider the following data set for a binary class problem

| A | B | Class Label |
|---|---|---|
| T | F | + |
| T | T | + |
| T | T | + |
| T | F | - |
| T | T | + |
| F | F | - |
| F | F | - |
| F | F | - |
| T | T | - |
| T | F | - |

Calculate the information gain when splitting on A and B where Class Label is the decision variable.
Out of A and B, which attribute would the decision tree induction algorithm choose?
c) Why recovery routines uses a log ?
In the process of recovery from a system failure, write short note on i) transaction-consistent check point; ii) transaction-oriented check point ; iii) indirect update and careful replacement.
(4M+3M+3M)

5.
a) Distinguish between dense and sparse index. When is it preferable to use a dense index rather than sparse index?

b) How distributed database are different from centralized databases? Discuss the utility of fragmentation in design and query processing of distributed databases. Why data replication is useful?

c) For a schedule consisting of a set of transactions, how deadlock can be detected? Explain with example. Discuss schemes for avoidance of deadlock.

d) Prove that two phase locking protocol ensures that the schedule is serializable.
(3M+4M+2M+1M)

---------- X ----------