Course No: 4411
Database Management Systems
Fall 2008 Midterm exam

Last Name:

First Name:

Student ID:

- Exam is 80 minutes.
- Open books/notes
- The exam is out of 20 points.

1. (16 points) **Multiple Choice** (sort of): if you pick the correct answer, no explanation is necessary. If you pick a wrong answer you may get a partial credit for an explanation.

   (a) Suppose 10 tuples of R can fit in one block, R has 100,000 tuples, and there are among these tuples 20,000 distinct tuples. We wish to eliminate duplicates in one pass. That is, we read blocks of R into a single buffer, one at a time, use B buffers (block-sized chunks of main memory) to hold intermediate results, and write the answer somewhere (it is not important where the output goes). The smallest possible value of B is:
   (a) 1000
   (b) 2000
   (c) 5000
   (d)* 10,000
   (e) 20,000

   (b) We wish to join the relations R(a; b), S(b; c), and T (c; d) under the following assumptions:

   1. There are B buffers available to hold blocks of data from these three relations.
   2. The relations occupy n_R, n_S , and n_T blocks, respectively.
   3. R is stored sorted by b; the other relations are unsorted.
   4. The particular strategy we shall use to perform the join is:

   i) Perform the first phase of two-phase multiway sort-merge on S. That is, as many times as necessary, load all buffers from S, sort the tuples on b, and write out the sorted sublist.

   ii) Load T entirely into main memory, using as many buffers as needed.

   iii) Merge (and join when appropriate) R and the sorted sublists of S, and compare each of the resulting tuples with the tuples of T . Any join tuple in the result is stored in an output buffer, not counted among the B buffers available for this process.

   Which of the following inequalities is the closest approximation to the condition under which this sequence of steps can be carried out as described?

   (a) * $B^2 > n_S + Bn_T$

(b) $B^2 > n_T + Bn_S$
(c) $B^2 > n_T + n_S$
(d) $B^2 > n_R + Bn_T + B^2 n_S$
(e) $B^2 > n_R + n_T + Bn_S$

(c) A B+ tree on a certain data file has three levels of nodes with an order equal 25 (that is, for the maximum of 50 search keys in a node). We assume it is maintained as described in class or the text; in particular, subminimum blocks due to deletion are merged and not allowed to exist. The minimum number of records in the data file is closest to which number:

(a) * 1350
(b) 14,400
(c) 15,000
(d) 15,625
(e) 120,050

(d) Relation R(a; b) is stored in 200 blocks, with 100 records (tuples) of R fitting in one block. Relation S(c; d) is stored in 100 blocks, with 50 records (tuples) per block. When we compute $R \bowtie_{b=c} S$, we find that the average tuple of R joins with five tuples of S. How many blocks does the result of the join require (you can assume the size of the tuple in the join is the union of the sizes of the input tuples)?

(a) 500
(b) 750
(c) 1000
(d) 2000
(e) * 3000

(e) Suppose relation R occupies N and relation S occupies M blocks. We assume that N $\leq$ M . Initially both relations are on disk, and we want to perform a nested-loop join in which we read each block of R and S only once (and do no other disk I/O other than writing the result). The minimum number of main-memory buffers (each the size of a block) that we need is:

(a) * N + 2
(b) M
(c) N + M - 1
(d) N + M
(e) NM

For the next two questions the following information is available on tables Sailors and Reserves.

- Reserves: 10,000 records
- Reserves.bid: 50 values

- Sailors: 1000 records
- Sailors.level: 10 values (1..10)

The primary key of Sailors is sid; of Reserves is sid+bid+day. Table Reserves has a foreign key on sid referencing Sailors (on sid). All columns are *not null*.

(f)
```
select distinct S.sid, R.day
    from Sailor S, Reserves R
    where S.sid = R.sid and
        R.bid = 7 and S.level = 6;
```

Estimate the selectivity of the above query as the number of tuples it likely returns.

(A) 2
(B) 5
(C) * 20
(D) 50
(E) 100
(F) 10,000

(g)
```
select distinct S.level
    from Sailor S, Reserves R
    where S.sid = R.sid and
        R.bid = 7 and S.level <= 6 and S.level >= 2;
```

Estimate the selectivity of the above query as the number of tuples it likely returns.

(A) 1
(B) * 5
(C) 20
(D) 80
(E) 100
(F) 10,000

(h) Consider table R with attributes A and B, table S with attributes B and C, and table T with attributes C and D. All joins below are natural joins; that is, the join columns are those columns named the same between the two tables.

I $\pi_B(R \bowtie (S \bowtie T))$
II $((\pi_B(R \bowtie S)) \bowtie T)$
III $((\pi_B(R) \bowtie S) \bowtie T)$

Which of the above relational algebra expressions necessarily evaluate to the same result?

(A) All three evaluate to the same result.
(B) * They each evaluate to different results.
(C) I and II
(D) I and III
(E) II and III
(F) Not enough information is provided to determine this.

2. **Short Questions** (4 points)

(a) In the execution of a block nested loops join of **R** and **S**, is a record of **R** ever compared with a record of **S** *and* they do *not* match (on the join attribute(s))? If not, briefly explain how the block nested loops join method avoids this. If so, give an example of how this can occur.

It does not happen.

(b) In the execution of a hash join of **R** and **S**, is a record of **R** ever compared with a record of **S** *and* they do *not* match (on the join attribute(s))? If not, briefly explain how the hash join method avoids this. If so, give an example of how this can occur.

It does not happen.