CS482 – Final

1:00 pm – 3:00 pm

December 9, 2004

Note:

- There are 5 questions (105 points: 100 points for 25% of the total grade and 5 bonus).
- For undergraduate students: you only need to get 95% to get full credit.
- Be short and precise in your answers.
- Merry Christmas

Name:

Signature:

1. (21 points) Assume that the rows of the table $R(a, b)$ are stored in an unsorted heap file and span over M pages, each page has C rows.

(a) If no index is available, how many I/O operations on average will be required to compute the result of the following statement:

```
SELECT * WHERE a = 100
```

Hint: You need to consider two cases: a is a key; a is not a key of R. Think about the number of rows in the answer.

(b) Assume that a has M possible values that are evenly distributed. Assume also that 100 is a possible value of a. What is the cost of computing the answer for the above query if we have a hash index on a? Assume that each bucket of the index fits in one page.

Remark: Here, you do not need to consider the two cases as in (a).

(c) Is the hash index in (b) a clustered index or a unclustered index?

2. (32 points) A table has 100,000 rows and each row occupies 200 bytes. The table is stored on a disk which has 4K bytes per page. Compute the maximum (worst case) cost (number of I/O operations) of doing an equality search on the primary key assuming the following access paths. Explain your answer.

(a) The data file is unsorted and has no index.

(b) The data file is sorted on the primary key and has no index.

(c) There is an unclustered hash index whose search key is the primary key. Assume all buckets are stored on disk and that each bucket has one overflow page.

(d) There is an unclustered B+ tree index whose search key is the primary key. Assume that each index entry in the tree occupies 20 bytes and the entire tree resides on the disk. Each page of the B+ tree uses 4000 bytes to store index entries.

3. (20 points) Consider a relation $Spouse$ over the attributes $SSN\_Husband$ and SSN_Wife with the following characteristics:

- 5,000 tuples with 10 tuples per page
- A 2-level B+ tree index on attribute $SSN\_Husband$ with up to 100 index entries per page
- Attribute $SSN\_Husband$ is a candidate key of S
- The values that the attribute $SSN\_Husband$ takes in relation $Spouse$ are uniformly distributed in the range 1 to 100,000.

Answer the following.

(a) Estimate the number of disk accesses needed to compute the range query

$$\sigma_{SSN\_Husband>1000 \wedge SSN\_Husband<6000}(Spouse)$$

if the index on $SSN\_Husband$ is unclustered.

Hint: First, estimate the number of records that the query will return. Next, think about the fan-out number and compute the pages in the leaf level of the tree.

(b) What would be the cost if the above index were clustered?

4. (20 points) Given two tables R and S with

- Table R occupies 800 pages, 20 rows per page, one of its attributes is A
- Table S occupies 200 pages, 10 rows per page, one of its attributes is B
- Main memory has 52 buffers

Answer the following.

(a) Compute the minimum cost (measured as the number of I/O operations) of a block-nested loops join $R \rhd\lhd_{A=B} S$. Explain your answer.

(b) Assume there is an unclustered, 3-level, B+ tree index on R with search key A and that each row of S joins with 5 rows of R. Compute the cost of an index-nested loops join. Provide the necessary justification.

5. (12 points) Consider a relation $Transcript$ over the attributes SSN, $Semester$, $Grade$, and $Crscode$ with the following characteristics:

- The relation has 100,000 rows, stored on 5,000 pages
- The main memory has 102 pages
- A 4-level B+ tree index on (SSN, CrsCode, Semester) with 500 pages at the leaf level
- A hash index on SSN stored on 200 pages
- A clustered index on (CrsCode, Semester) stored on 400 pages

Assume that the number of index entries in each index page is evenly distributed. What will you use to answer the following query? Justify your answer.

(a)

$$\sigma_{SSN=123 \wedge CrsCode='CS482' \wedge Semester='F03'}(Transcript)$$

(b)

$$\sigma_{CrsCode='CS482' \wedge Semester='F03' \wedge (Grade='A' \vee Grade='B')}(Transcript)$$