CSE 322 Midterm III

Olcay Taner YILDIZ

I. Question (14 points)

- What is the advantage of directory of pages compared to the link list of pages?
- How many records can you store in a DBMS page of size 4KB, where the record length is fixed (10 bytes) and you want to keep track of free slot information.

II. Question (12 points)

Consider a delete specified using an equality condition. For each of the four file organizations (heap file, sorted file, clustered tree index, clustered hash index), what is the cost if no record qualifies?

- The number of data pages when records are packed onto pages with no space is B.
- The number of records per page is R.
- The average time to read and write disk page is D.
- The fanout is F.

III. Question (18 points)

Suppose you have a file with 1024 pages and you have 5 buffer pages. Assuming that we use all buffer pages in sorting, answer the following questions:

- How many runs will you produce in the pass 0?
- How many passes will it take to sort the file completely?
- How many buffer pages do you need to sort the file completely in just three passes?

IV. Question (18 points)

What are the minimum and maximum number of nodes in a B+ tree of order 2 with height h.

V. Question (16 points)

Consider the following schema with the Sailors relation:

Sailors(sid:integer, sname:string, rating:integer, age:real)

For each of the following indexes, list whether the index matches the given selection conditions:

Index	Index fields	Condition
B+ tree index	sid	sid < 50000
B+ tree index	sid	sid = 50000
Hash index	sid	sid < 50000
Hash index	sid	sid = 50000
B+ tree index	sid, age	sid < 50000 and age = 21
B+ tree index	sid, age	sid = 50000 and age = 21
Hash index	sid, age	age = 21
Hash index	sid, age	sid = 50000 and age = 21

VI. Question (22 points)

- Give an example of Extendible Hashing where inserting an entry increases the global depth.
- Give an example of Extendible Hashing where the global depth is 4 and there are buckets of local depths 1, 2, 3 and 4.

VII. Bonus Question (30 points)

Consider the following schema:

Suppliers(sid:integer, sname:string, rating:integer, address:string)
Catalog(sid:integer, pid:integer, cost:real)

There is an hash index on pid of the Catalog table, a B+ tree index on rating of the Suppliers table and a primary hash index on sid of the Suppliers table. Rating field can take values between 1 and 10. pid field can take values between 1 and 100. You can assume the field values are uniformly distributed. Assume that there are 100000 Suppliers tuples in 1000 pages. Similarly, there are 40000 Catalog tuples in 500 pages. Given the following query:

```
SELECT S.sname
FROM Suppliers S, Catalog C
WHERE S.sid = C.sid AND
S.rating < 4 AND C.pid = 50
```

This query can be expressed in relational algebra in three different forms:

$$\pi_{sname}(\sigma_{rating<4 \wedge pid=50}(Suppliers \bowtie Catalog)) \tag{1}$$

$$\pi_{sname}((\sigma_{rating<4}Suppliers) \bowtie (\sigma_{pid=50}Catalog)) \tag{2}$$

$$\pi_{sname}(\sigma_{rating<4}((\sigma_{pid=50}Catalog) \bowtie Suppliers)) \tag{3}$$

For each form the query plan is as follows:

1) File scan Catalog and Suppliers tables, join two tables using simple nested loops, do the selection and projection on the fly.
2) File scan Suppliers table make the selection and put the result in a temporary table. File scan Catalog table make the selection and put the result in a temporary table. Join the results using Sort-merge-join and make the projection on the fly.
3) Use hash index on pid of Catalog table to make the selection, do not write the result to a temporary table, use hash index on sid of the Suppliers table to make join using index nested loops. Do the last selection and projection on the fly.

What is the cost of each plan in terms of I/O costs?