Fall 2004 CS 186 Exercise Questions

Week 8 – Ending 10/22

QUERY OPTIMIZATION

1) Consider the following relational schema and SQL query:

```
Emp(eid,did,sal,hobby)
Dept(did,dname,floor,phone)
Finance(did,budget,sales,expences)
```

```
SELECT D.dname, F.budget
FROM Emp E, Dept D, Finance F
WHERE E.did = D.did AND D.did = F.did AND
D.floor = 1 AND E.sal >= 59000 AND E.hobby = ‘yodelling’
```

a) Identify a relational algebra expression that reflects the order of operations that a decent query optimizer would choose.

b) List the join orders that a System R query optimizer would consider. (Assume that the optimizer follows the heuristic of never considering plans that require the computation of cross-products).

c) Suppose that the following additional information is available:

- Unclustered B+ tree indexes exist on Emp.did, Emp.sal, Dept.did, and Finance.did (each leaf page contains up to 200 entries).
- The system’s statistics indicate that employee salaries range from 10,000 to 60,000, employees enjoy 200 different hobbies
- The company owns two floors in the building.
- There are a total of 50,000 employees and 5,000 departments (each with corresponding financial information) in the database.
- The DBMS used by the company has just one join method available, namely index nested loops.

i) For each of the query’s base relations estimate the number of tuples that would be initially selected from that relation if all of the non-join predicates on that relation were applied to it before any join processing begins.

ii) Given your answer to the preceding question, which of the join orders that are considered by the optimizer has the least estimated cost?

2) You have a database with the following schema:

```
Employee (name, address, phone, dept)
Department (name, floor, manager, dept, area, budget)
Contracts (company, contact, area, rating, state)
```

- E.dept is a foreign key to D.dept
- C.area is a foreign key to D.area
- D.budget is between $5000 and $30000
- C.state has 50 unique values
- C.rating is between 1 and 10

- There are 10,000 employees, 10 per page, 1,000 pages
- There are 50 departments, 10 per page, 5 pages
- There are 1,000,000 contracts, 100 per page, 10,000 pages

- For B-Tree indexes, there are 100 keys per node.
- Unclustered B+Tree Index on E.dept, 50 unique values
- Unclustered B+Tree Index on D.dept, 50 unique values
- Unclustered Hash Index on C.company, 2,000 unique values
- Clustered B+Tree Index on C.state, C.rating

Query 1

```
SELECT *
FROM employee E, department D
WHERE E.dept = D.dept AND D.budget > $10000
```

1) Begin the process of query optimization, by determining all the cheapest and interesting access methods for each relation and their costs. (Pass 1)

2) Enumerate all the two-way joins that the optimizer should estimate costs for, you can use nested loops, index nested loops, hash, and sort-merge join algorithms. You do not need to estimate the costs for each plan

Query 2

```
SELECT *
FROM employee E, department D, contracts C
WHERE E.dept = D.dept AND D.area = C.area AND
D.budget > $10000 AND C.state = CA AND C.rating > 5
```

1) Enumerate and estimate costs for all plans in Pass 1
2) Enumerate plans considered in Pass 2