CPSC 404: Advanced Relational Databases
Quiz 2
Laks V.S. Lakshmanan

Date: Wednesday, November 12, 2003. **Duration: 45 minutes**.

IMPORTANT INSTRUCTIONS:

- Everyone must copy the following honor code verbiage in their answer booklet and sign: "I am aware of what constitutes academic misconduct and the disciplinary actions that may be taken against it, and agree not to cheat."
- Exams without a signed honor code will not be marked and the student will be presumed to have missed the quiz effectively.
- Be sure to write your name in block letters and print your student ID.
- Answer **all** questions.
- Show all your work, including all the steps. State your assumptions, if any, clearly. Answers without proper explanation or details will get a low credit.
- At the end of the quiz, you must hand in *both* your answer booklet and this quiz (i.e., the questions).
- The term *cost* in all questions refers to I/O cost.
- The budgeting accounts for a 5 min. reviewing of answers.
- There is a bonus question (# 4) that is worth 10% (question not shown in this sample).

Question 1 [*40%; Suggested Time:15 min.*]
Consider the relation `Employee(ename, title, dname, salary)`, containing 2 million tuples, each 100 bytes long. The size of each page is 4K and there are 500 buffer pages. The relation contains duplicates. Consider the query:

```
SELECT DISTINCT *
FROM Employee
```

Outline how you would implement this query using hashing, indicating which attributes you'd hash on. Your strategy must be as efficient as possible, in terms of disk I/Os. Explain how many partitions you would produce in your strategy and why. Then estimate the cost of your strategy, showing your estimate for each step of the strategy.

Question 2 [*20%; Suggested Time:10 min.*]
Consider the above relation `Employee(ename, title, dname, salary)`. Suppose it contains 8 distinct values in column `title`. For simplicity, assume the 8 `title`-values are represented as numbers 1-8. Answer the following questions.

(a) Estimate the reduction factor of the condition `title` $>$ 3. What assumption (if any) did you make to obtain this estimate?

(b) Suppose the following frequency table is given to you, showing how frequently each of the 8 `title`-values appears in the `Employee` relation. Estimate the reduction factor using an equi-width histogram. Use 4 buckets.

Value	Frequency	Value	Frequency
1	500,000	5	550,000
2	50,000	6	350,000
3	350,000	7	25,000
4	50,000	8	225,000

(c) Using the same frequency table, now estimate the reduction factor using an equi-depth histogram, again using 4 buckets.

Question 3 [*40%; Suggested Time:15 min.*]
Relation Emp($\underline{\texttt{ename}}$, `title, dname, salary`) contains 200,000 tuples with 20 tuples/page. Relation Dept($\underline{\texttt{dname}}$, `dlocation, boss`) contains 40,000 tuples with 20 tuples/page. The primary key of each relation is underlined. Both relations are unsorted and have no index. Available buffer size is 500 pages. We wish to compute `Emp` $\bowtie_{\texttt{Emp.dname=Dept.dname}}$ `Dept`.

(a) Briefly outline a block-based (i.e., chunk-based) nested loops algorithm for computing the above join, indicating the chunk size. Your algorithm must be as efficient as possible. Estimate the cost of computing the join using your algorithm. Which relation would you put in the inner loop, and why? Show all your work, step by step.

(b) Suppose you wish to compute this join using sort-merge. What is the minimum no. of I/Os this strategy would cost? Do you have enough buffer pages to make this work? Explain.

All the best!