CPS216 Advanced Database Systems - Fall 2008
Assignment 2B

- Due date: Thursday, Oct. 16, 2008, in class (2.50 PM). Late submissions will not be accepted.
- Submission: In class, or email solution in pdf or plain text to shivnath@cs.duke.edu.
- Do not forget to indicate your name on your submission.
- State all assumptions. For questions where descriptive solutions are required, you will be graded both on the correctness and clarity of your reasoning.
- Email questions to shivnath@cs.duke.edu.

Question 1 **Points 15**

The following information is available about relations R and S:

- Relation R is clustered and the blocks of R are laid out contiguously on disk. B(R) = 1250 and T(R) = 12,500.
- Relation S is clustered and the blocks of S are laid out contiguously on disk. B(S) = 1000 and T(S) = 10000.
- M = 101 blocks.

a. For this question assume that our cost model is the same as the one we have been using in class, namely, the total number of blocks read or written, excluding the writes for the final output. Compute the number of buckets and the cost for the most efficient Hybrid Hash Join of relations R and S.

b. Suppose everything in the question remains the same except now M=51. Compute the number of buckets and the cost for the most efficient Hybrid Hash Join of relations R and S.

Question 2 **Points 5**

State two disadvantages of the following cost formula used in the Selinger paper:
COST = PAGE_FETCHES + W * (RSI CALLS)

Question 3 **Points 15**

We want to execute the following query:

```
Select R.B, S.C
From   R, S, T
Where R.A = 10 and R.B = S.B and S.C = T.C
```

Suppose we have the following statistics:

- $T(R) = 1000$, $T(S) = 2000$, $T(T) = 3000$
- $V(R, A) = 100$
- $V(R, B) = 100$
- $V(S, B) = 200$
- $V(S, C) = 500$
- $V(T, C) = 100$

1. If we use the calculations as per the Selinger paper, how many tuples will be returned in the result of this query?
2. As per the Selinger paper, what is the value of RSICARD (the number of expected RSI calls) for this query?

Question 4 **Points 15**

1. See the cost formula for tuple nested-loop join and the join by merging scans on Page 29 of the Selinger paper. These two formula are the same, which seems contrary to what we learned about in class. What do you think? Explain your thoughts briefly.
2. Suppose we have the following information:
   - Relation R is clustered, and the blocks of R are laid out contiguously on disk in order of attribute A. B(R) = 50 and T(R) = 1000.
   - Relation S is clustered, and the blocks of S are laid out contiguously on disk in order of attribute A. B(S) = 100 and T(S) = 1000.
   - M = 101 blocks.

   For this question assume that our cost model is the same as the one we have been using in class, namely, the total number of blocks read or written, excluding the writes for the final output. What is the cost of a tuple nested-loop join on R.A = S.A with R as the outer? What is the cost of a merging scans join on R.A = S.A?