

Exercises and hand-in

Advanced database technology

March 25, 2004

Suppose we have relations R_1 , R_2 and R_3 with common attributes A (appearing in R_1 and R_2) and B (appearing in R_2 and R_3). Tuples in all relations have fixed length, 5 tuples per disk block, and attributes A and B occupy 10% of the total tuple length. The relations occupy $B(R_1) = 8.000$, $B(R_2) = 10.000$ and $B(R_3) = 90.000$ blocks on disk, respectively. Values in attribute A are uniformly distributed in the domain $\{1, \dots, 400\}$, and values in attribute B are uniformly distributed in $\{1, \dots, 600\}$. Consider the following relational algebra expression:

$$\delta(\pi_A((\sigma_{A>200}(R_1)) \bowtie (\sigma_{B=4}(R_3)) \bowtie (\sigma_{A\leq 450}(R_2)))).$$

1. Estimate the sizes of all subexpressions using the formulas from the lecture. (Ignore that you might be able to make a better estimate!)
2. Using these estimates, apply dynamic programming (Selinger-Style Optimization, see G UW page 845) to find the best physical query plan *not* using any indexes:
 - Determine the order of joins.
 - Determine the algorithms used for all operations. Assume that there is memory for either a two-pass sorting based join using $5(B(R_i) + B(R_j))$ I/Os to join R_i and R_j , or a two-pass hash join using $3(B(R_i) + B(R_j))$ I/Os to join R_i and R_j .
 - Determine where to use pipelining. Assume that there are 10 extra memory buffers available for pipelining purposes.

The following exercise is to be handed in at the latest April 15 at 10.00 AM.

Problem 4 from the ADBT exam, June 2003. (See News section on home page.)

Hand in before the lecture on that day, or earlier in our mail boxes in the information office.