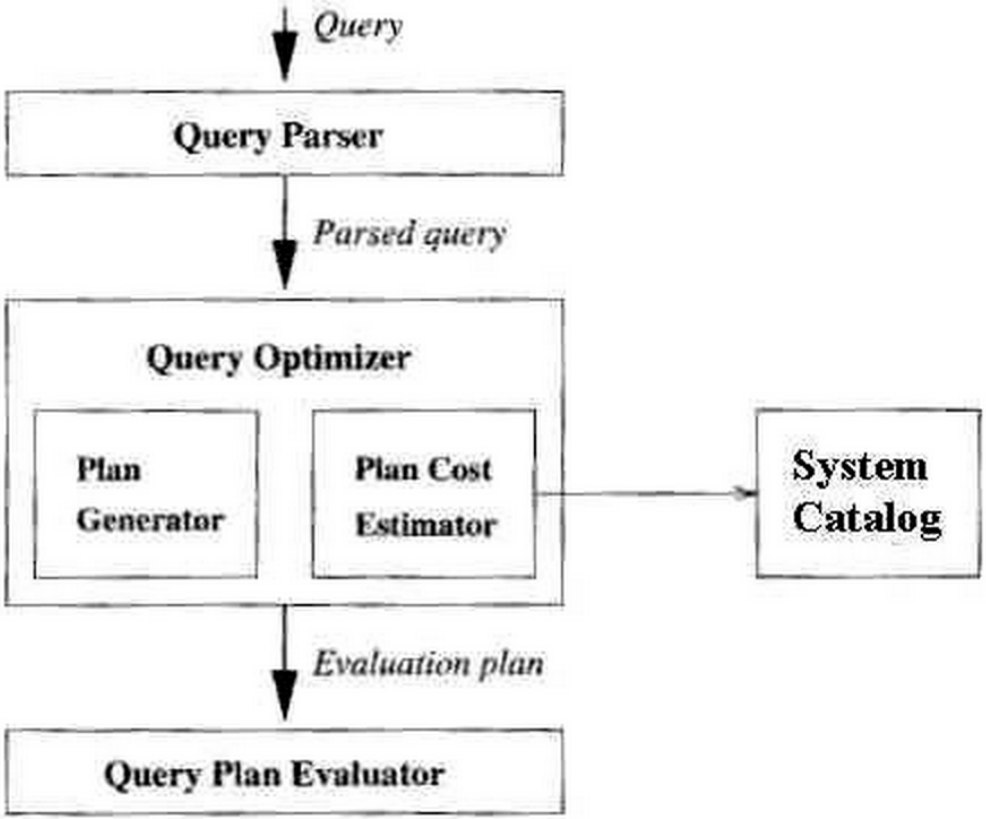
HELLO!

Rate-Based Query Optimization for Streaming Information Sources

Stratis D. Viglas

Jeffrey F. Naughton


Query
Query Parser
Parsed query
Query Optimizer
Plan Generator
Plan Cost Estimator
System Catalog
Evaluation plan
Query Plan Evaluator

Cardinality Based vs. Rate Based

Cost Estimation

Let us consider two select operations A and B. Assume that the selectivity for A is 0.1 and B is 0.2 and that the input size is 500.

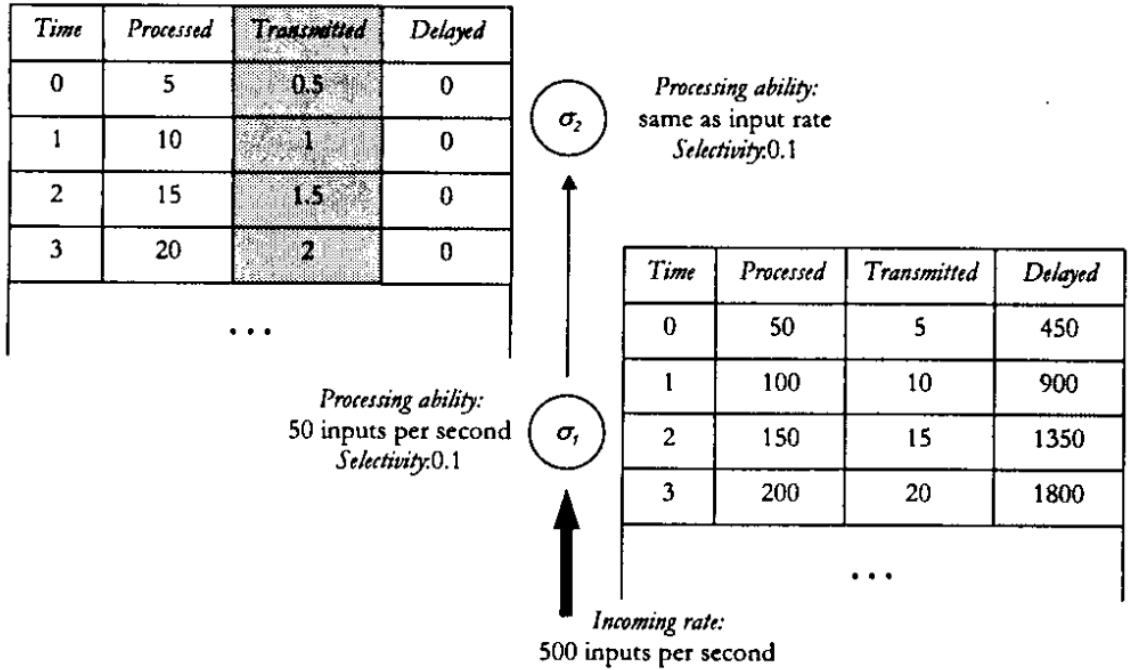
$$\text{Cost}(A \rightarrow B) = 500 * c_{A} + 500 * 0.1 * c_{B}$$

$$\text{Cost}(B \rightarrow A) = 500 * c_{B} + 500 * 0.2 * c_{A}$$

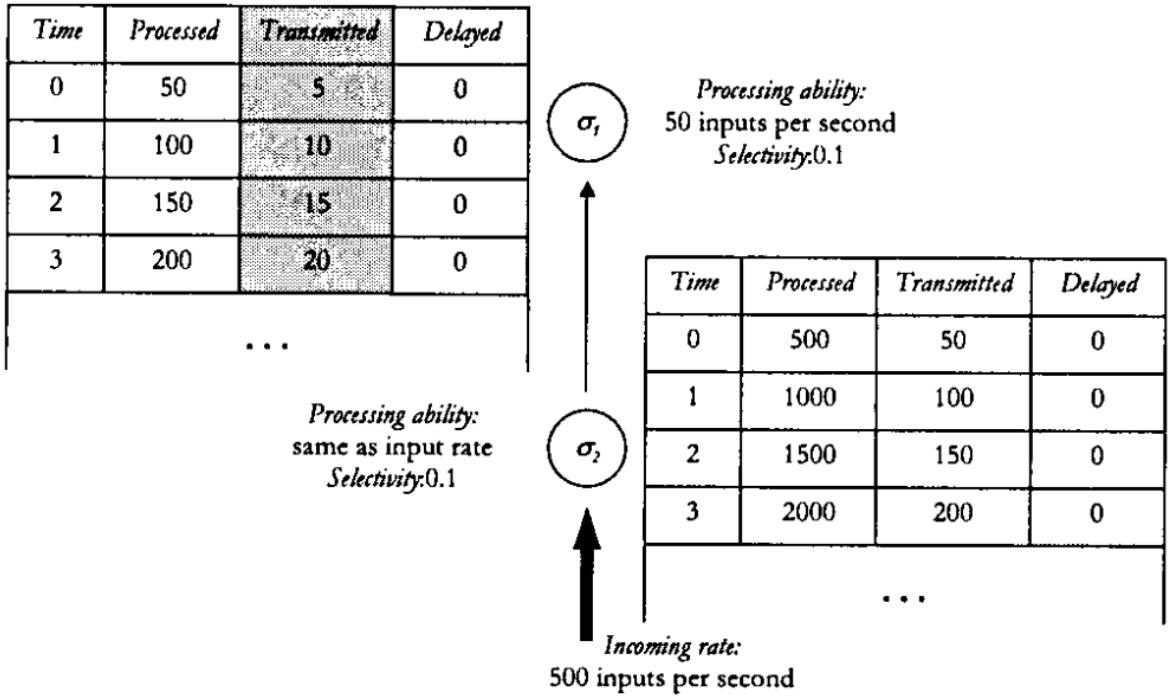
Assume that the selectivity of each of A and B is 0.1; input arrives at 500 tuples per second; A can process 50 inputs per second and B can process data as fast as it receives it.

Size of input is infinite

$\Rightarrow$ Cost of each plan is infinite



(a) Output rate = 0.5 outputs per second



(b) Output rate = 5 outputs per second

~~*“What is the cost of this query plan?”*~~

“What is the expected output rate of this query plan?”

$$Output\ rate = \frac{Number\ of\ outputs\ transmitted}{Time\ needed\ to\ make\ the\ transmission}$$

Table 1: Cost variables used in the estimation of output rates

Cost Variable	*Meaning*
C_π	Cost of projecting parts of an input object
C_σ	Cost of performing a selection on an input object
C_l	Cost of handling an input coming from the left-hand side of a join
C_r	Cost of handling an input coming from the right-hand side of a join
T	Cost of making a single transmission

r_o Output Rate

r_i Input Rate

r_r Right Input Rate

r_l Left Input Rate

Projections

$$r_o = r_i$$

Selections

$$r_o = f \cdot r_i$$

Joins

$$r_o = \frac{f \cdot r_l \cdot r_r \cdot t}{r_l \cdot C_l + r_r \cdot C_r}$$

Optimize for a specific time point in the execution process using local rate maximization

Optimize for output production size using local time minimization

Experimental Validation

Rate Based Cost Model

Does the cost model correctly estimate individual plan performance?

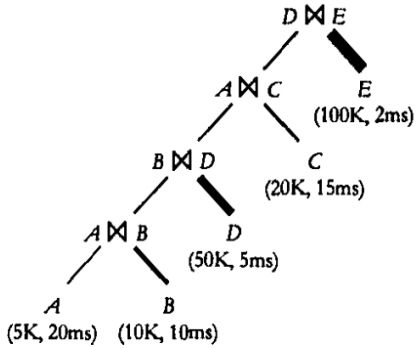
Is the framework capable of providing correct decisions regarding the best choice among a set of plans?

5 XML data sources

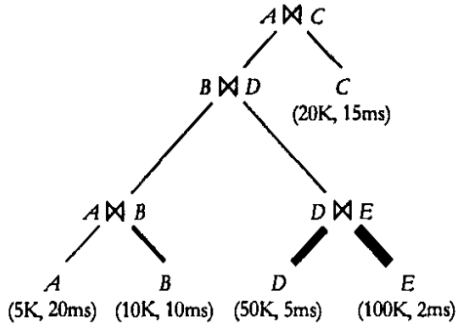
Wide range of selectivities

Source	*Number of tuples*	*Size*
A	5,000	0.7 MB
B	10,000	1.5 MB
C	20,000	1.8 MB
D	50,000	5.9 MB
E	100,000	9.3 MB

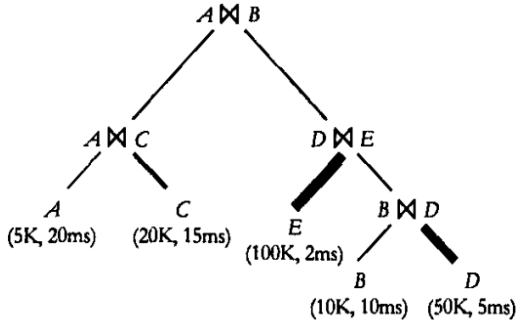
5 Way Equi Join



(a) Left Deep



(b) Fast Leaves



(c) Evenly Spread

Comparison to Traditional Cost Model

Plan	*Traditional Estimation*	*Rate-Based estimation*
Left Deep	10^4	$1.3 \cdot 10^3$
Fast Leaves	$2 \cdot 10^3$	$9.7 \cdot 10^2$
Evenly Spread	$5 \cdot 10^3$	$8.8 \cdot 10^2$

Rate Based Estimation is the way to go!

THANKS!

Any questions?