Anorexic Plan Diagrams

E0 261

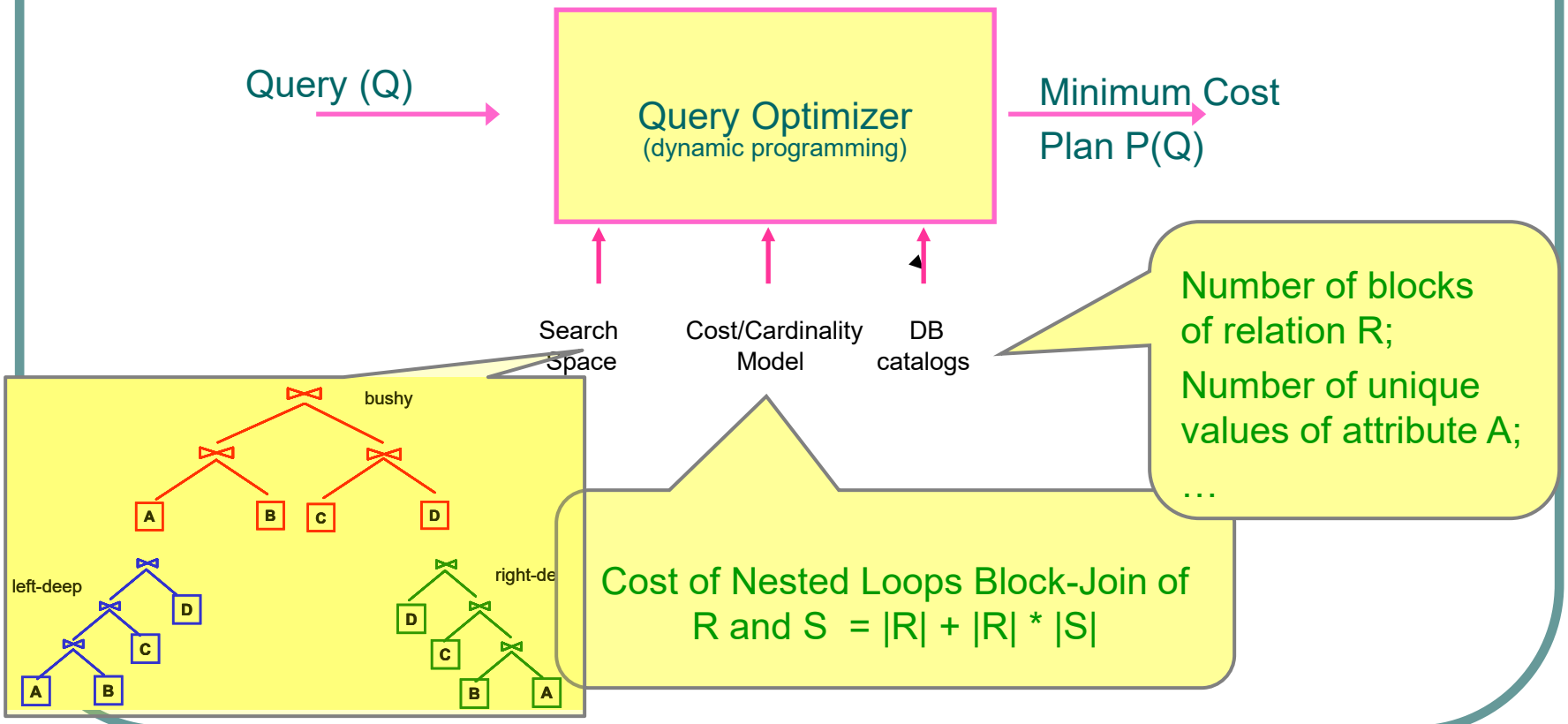
Jayant Haritsa

Computer Science and Automation

Indian Institute of Science

Query Plan Selection

- Core technique



Relational Selectivities

- Cost-based Query Optimizer's choice of

  execution plan = f (query, database, system, …)

- For a given database and system setup,

  execution plan = f (selectivities of query's base relations)

  – selectivity is the estimated percentage of rows of a relation used in producing the query result (i.e. normalized cardinality)

Query Template [Q7 of TPC-H]

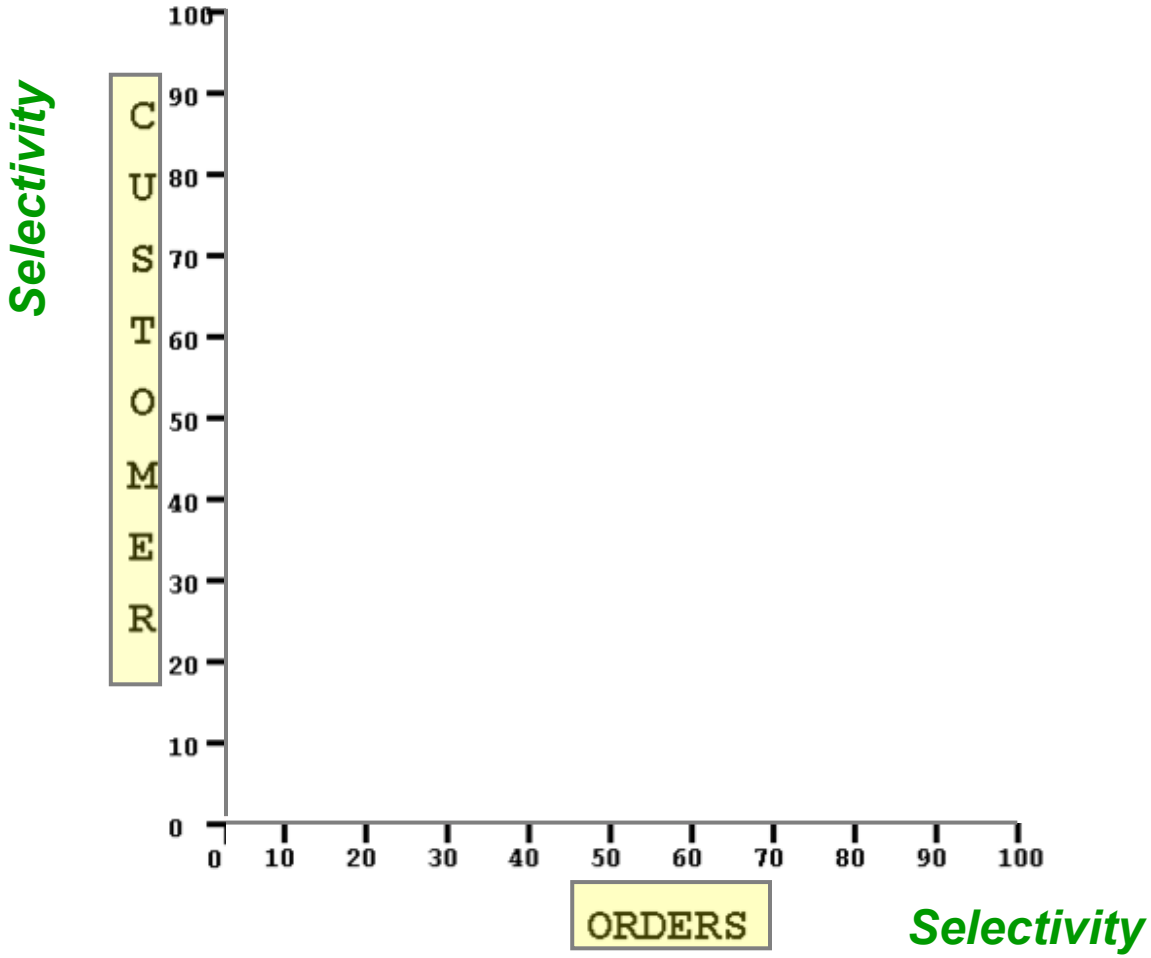
Determines the values of goods shipped between nations in a time period

```
select
    supp_nation, cust_nation, l_year, sum(volume) as revenue
from
(select n1.n_name as supp_nation,  n2.n_name as cust_nation,
        extract(year from l_shipdate) as l_year,
        l_extendedprice * (1 - l_discount) as volume
 from   supplier, lineitem, orders, customer, nation n1, nation n2
 wher                    okey and o_orderkey = l
                         key and s_nationkey = n
                         n2.n_nationkey and
        ((n1.n_name =  RANCE' and n2.n_name = 'GE     NY') or
         (n1.n_name = 'GERMANY' and n2.n_name = '     RANCE')) and
         l_shipdate between date '1995-01-01' and da   '1996-12-31'
         and  o_totalprice ≤ C1  and  c_acctbal ≤ C2   )  as shipping
 group by supp_nation, cust_nation, l_year
 order by supp_nation, cust_nation, l_year
```

Value determines selectivity of ORDERS relation

Value determines selectivity of CUSTOMER relation

Relational Selectivity Space

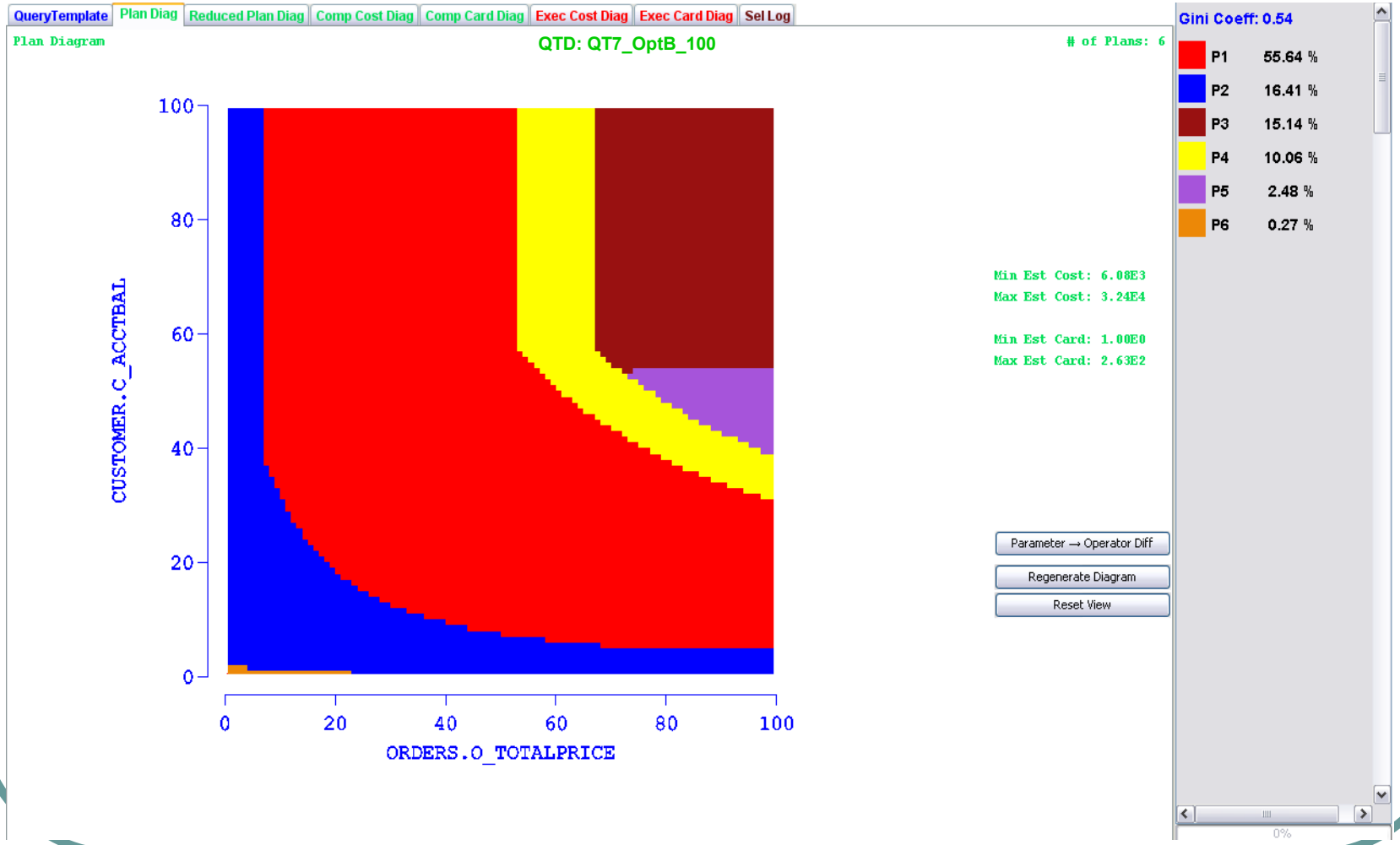


Plan, Cost, Card Diagrams

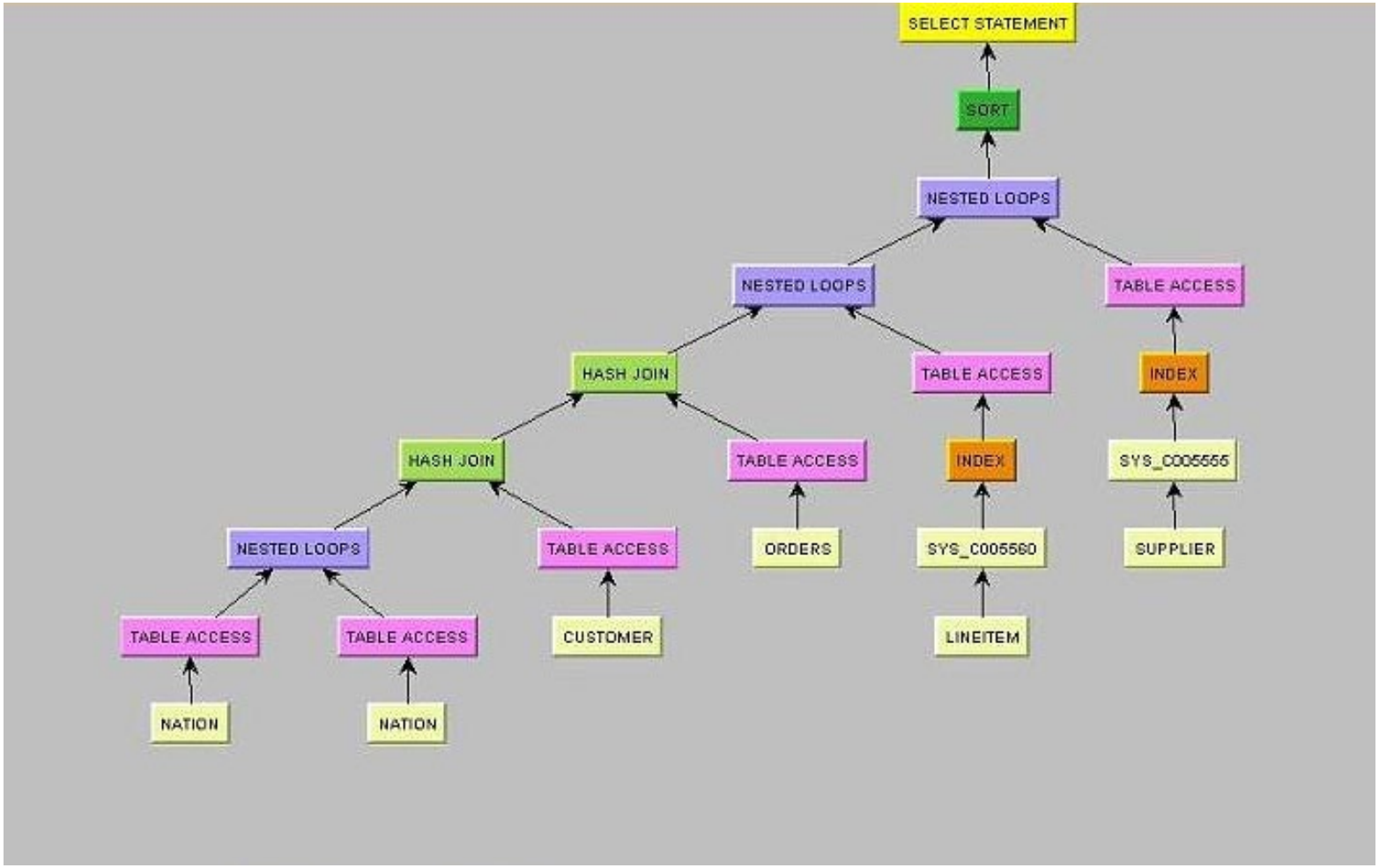
- A plan diagram is a pictorial enumeration of the plan choices of the query optimizer over the relational selectivity space
- A cost diagram is a visualization of the (estimated) plan execution costs over the same relational selectivity space
- A card diagram is a visualization of the (estimated) query result cardinalities over the same relational selectivity space

Sample Plan Diagram

[QT7, OptB, Res=100]

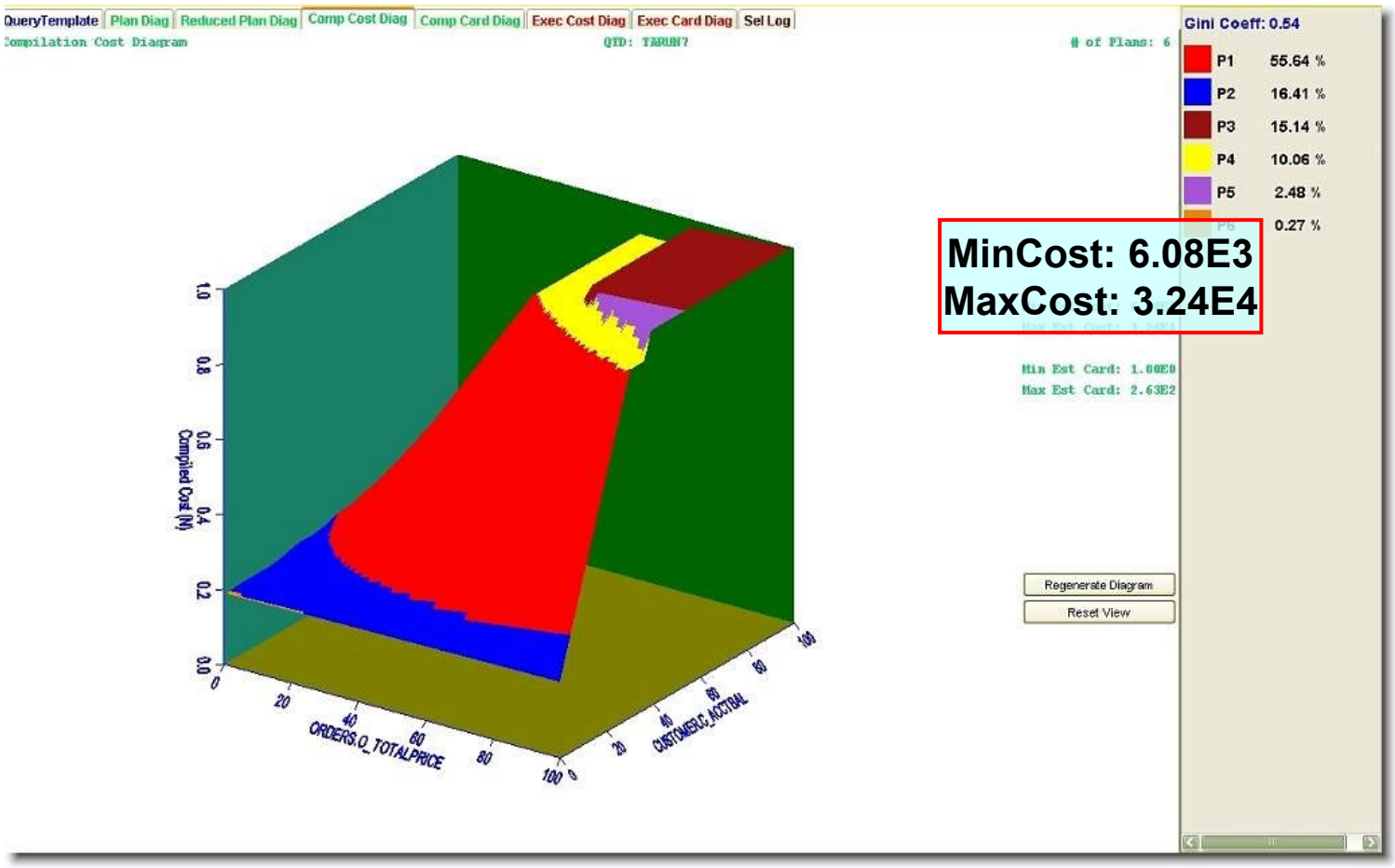


Plan P3


SELECT STATEMENT
SORT
NESTED LOOPS
NESTED LOOPS
TABLE ACCESS
HASH JOIN
TABLE ACCESS
INDEX
HASH JOIN
TABLE ACCESS
INDEX
SYS_C005555
NESTED LOOPS
TABLE ACCESS
ORDERS
SYS_C005560
SUPPLIER
TABLE ACCESS
TABLE ACCESS
CUSTOMER
LINEITEM
NATION
NATION

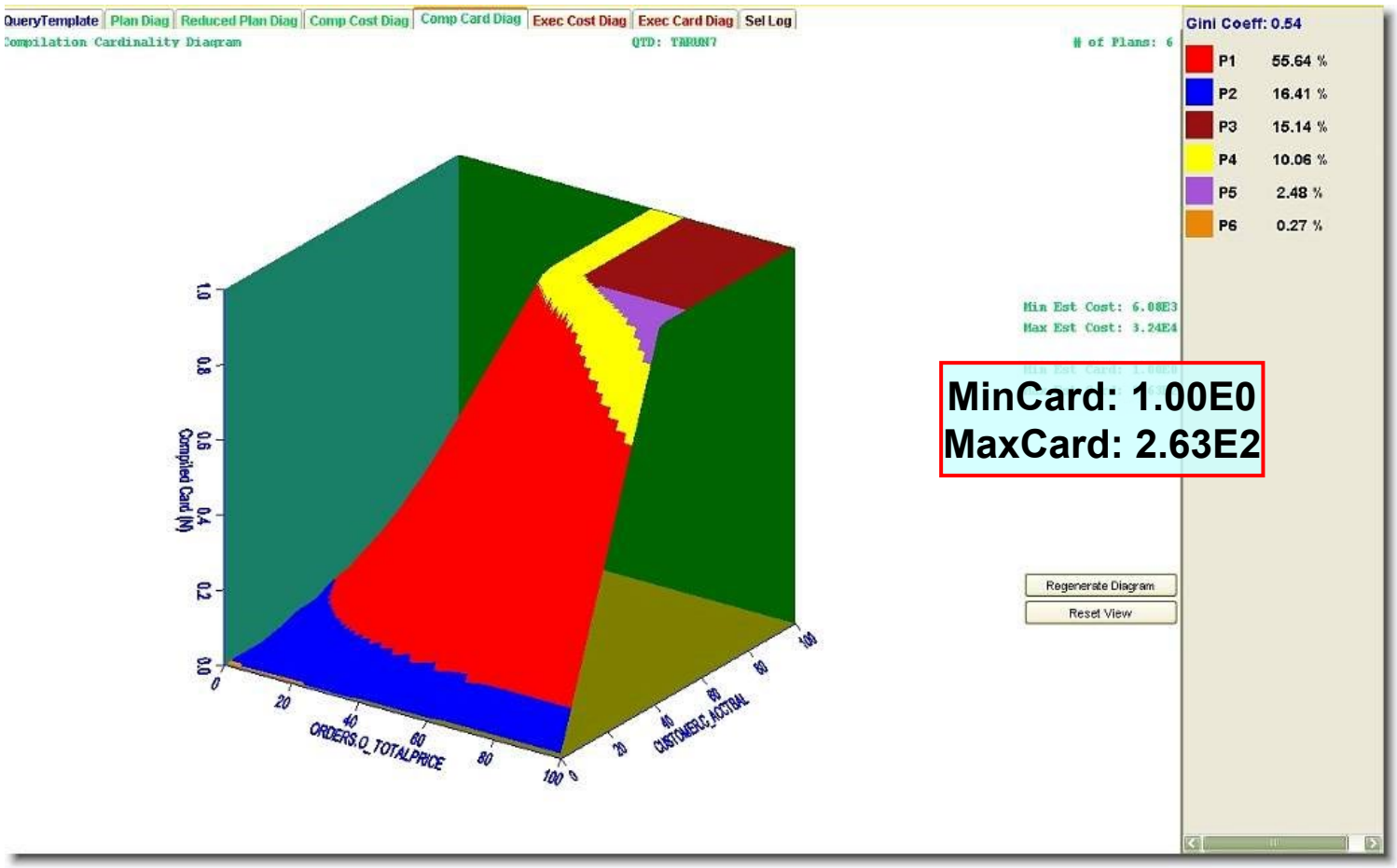
Sample Cost Diagram

[QT7,OptB]



Sample Cardinality Diagram

[QT7,OptB]



Part I: PICASSO [VLDB05]

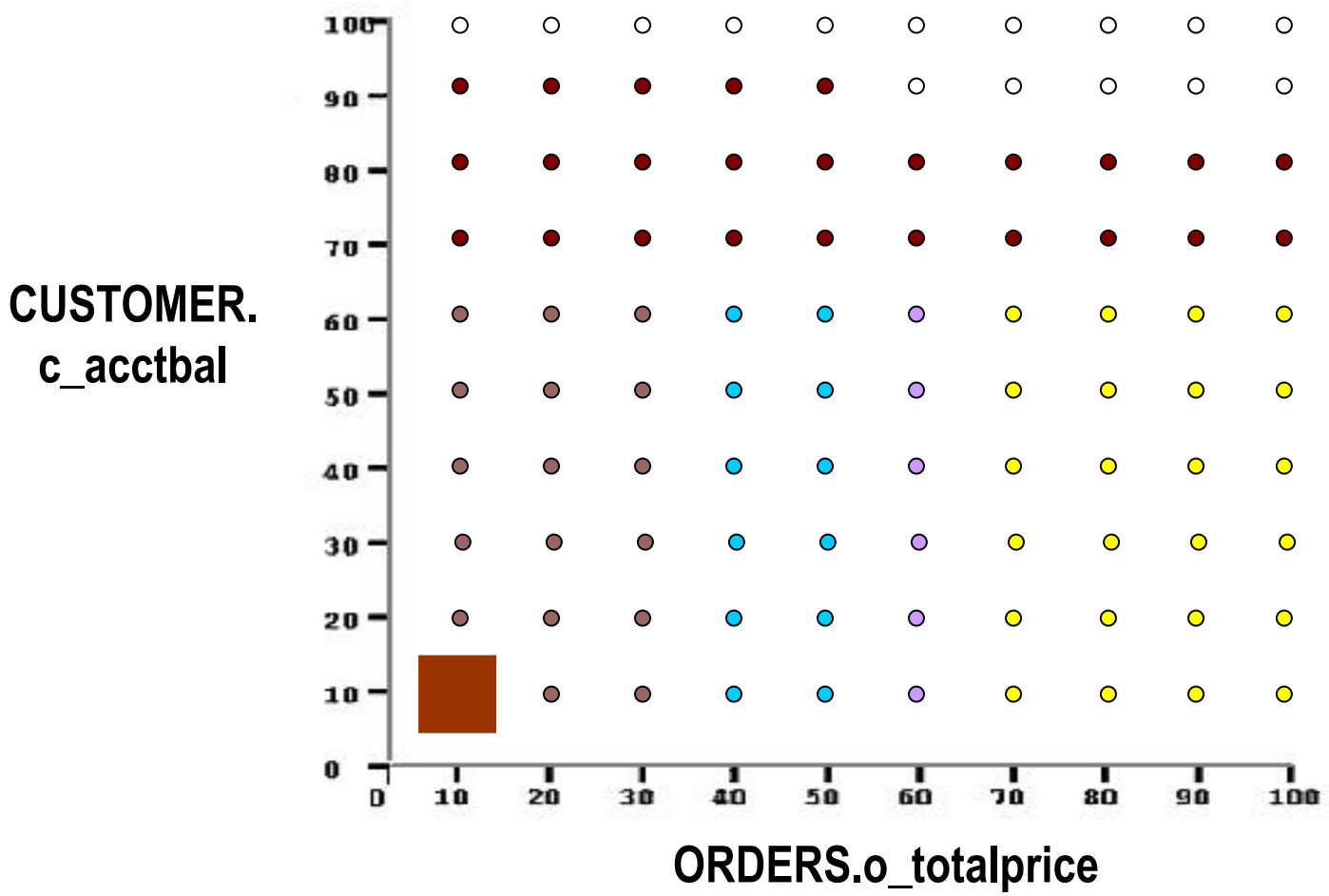
Overview

Picasso is a Java tool that, given an *n*-dimensional SQL query template and a choice of database engine, automatically generates plan, cost and card diagrams

- Fires queries at user-specified granularity (10, 30, 100, 300, 1000 queries per dimension)
- Visualization: 2D plan diagrams (slices if n > 2)
  3D cost and card diagrams

Also:
- Plan-trees, Plan differences
- Execution cost/card diagrams
- Abstract-plan diagrams
- Foreign Engine Plans

Diagram Generation Process



The Picasso Connection

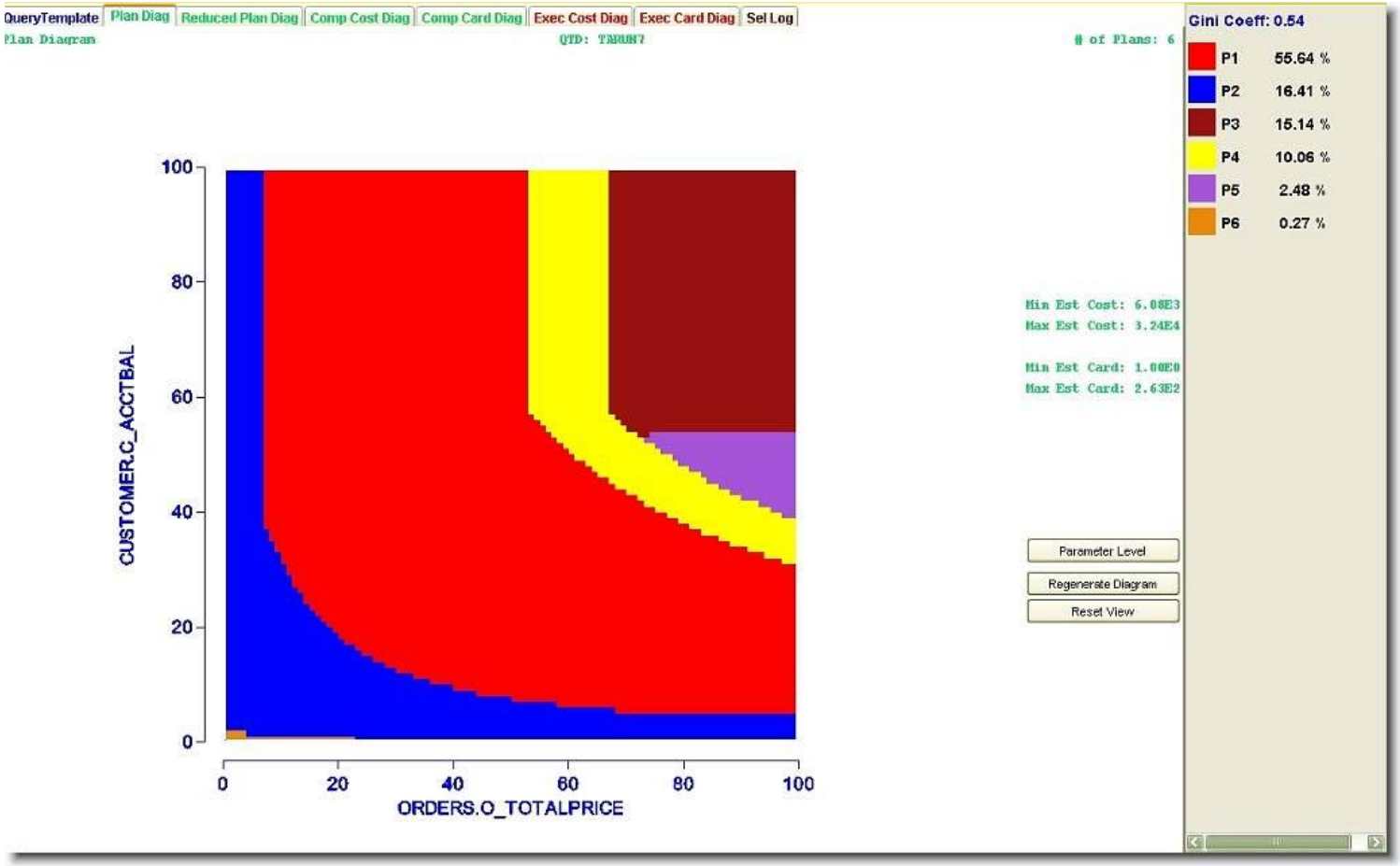
Plan diagrams are often similar to cubist paintings !

[Pablo Picasso – founder of cubist genre]

Woman with a guitar

Georges Braque, 1913

Smooth Plan Diagram
[QT7,OptB]

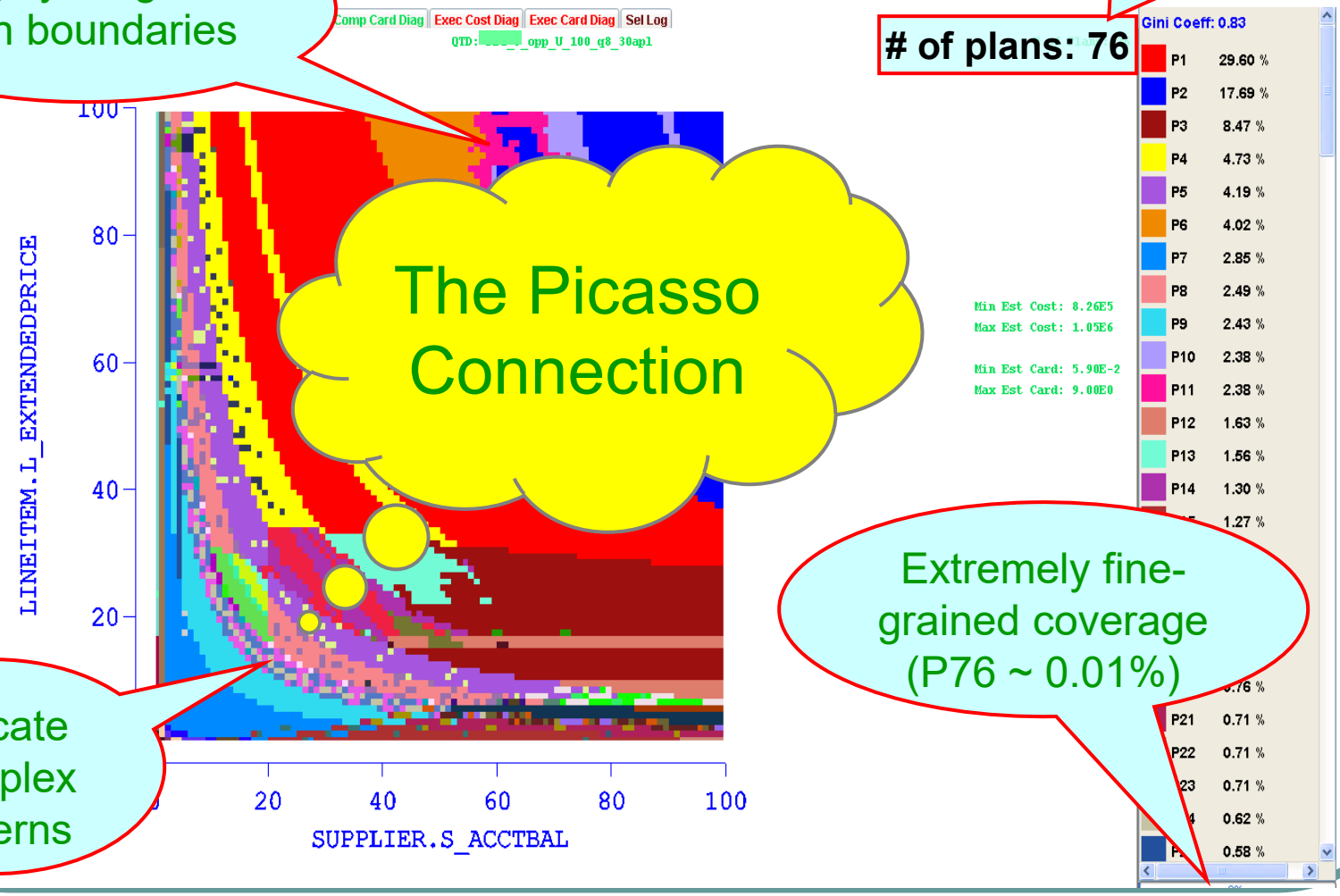


Complex Plan Diagram

[QT8.OptA*]

Highly irregular plan boundaries

Increases to **90 plans** with 300x300 grid !

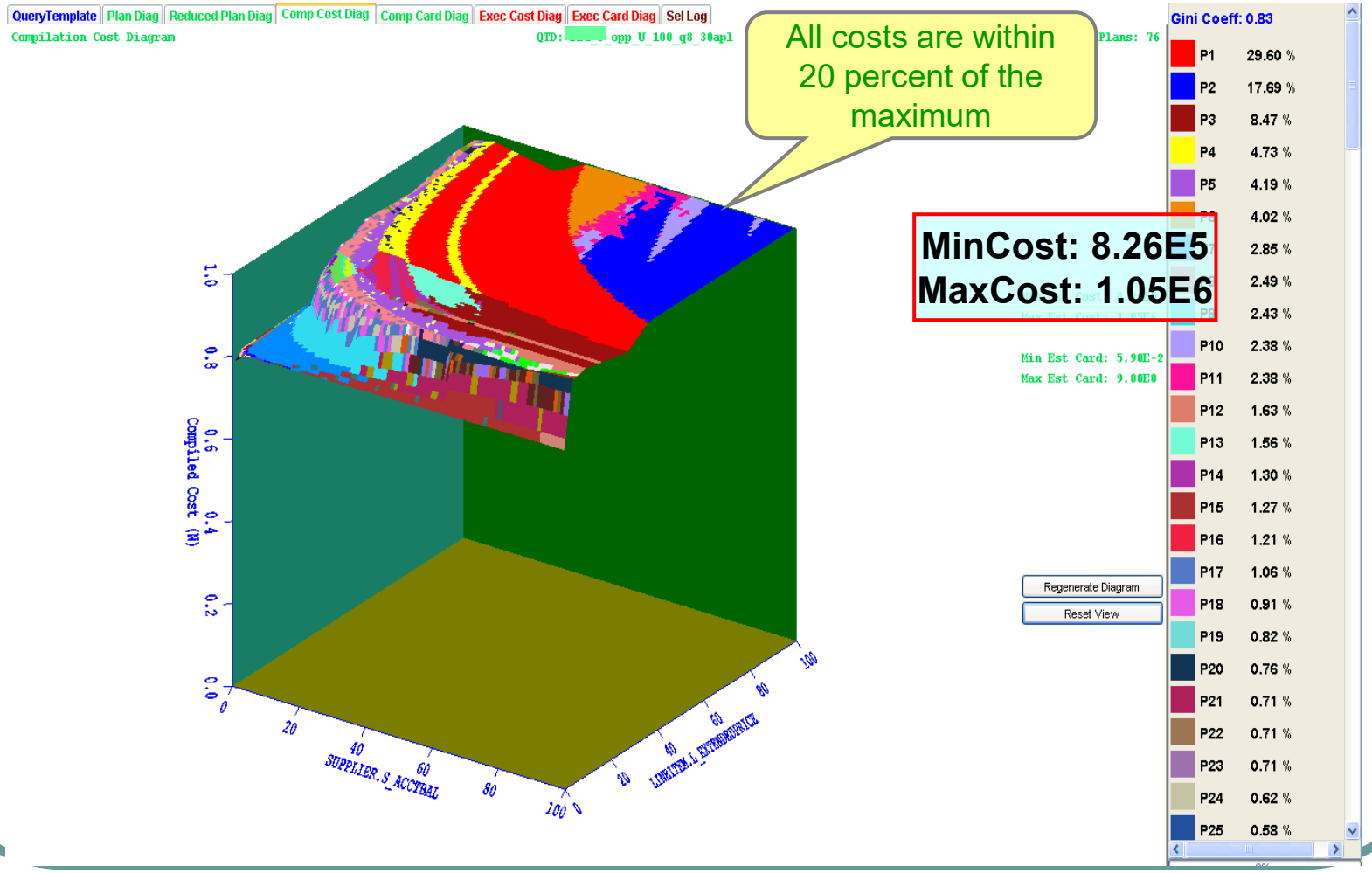


The Picasso Connection

Extremely fine-grained coverage (P76 ~ 0.01%)

Intricate Complex Patterns

Cost Diagram
[QT8, Opt A*]



Remarks

- Modern optimizers tend to make extremely fine-grained and skewed choices
- Is this an over-kill, perhaps not merited by the coarseness of the underlying cost space – i.e. are optimizers "doing too good a job" ?
- Is it feasible to reduce the plan diagram complexity without materially affecting the query processing quality?

Part II: PLAN DIAGRAM REDUCTION [VLDB07]

Problem Statement

Can the plan diagram be recolored with a smaller set of colors (i.e. some plans are "swallowed" by others), such that

Guarantee:

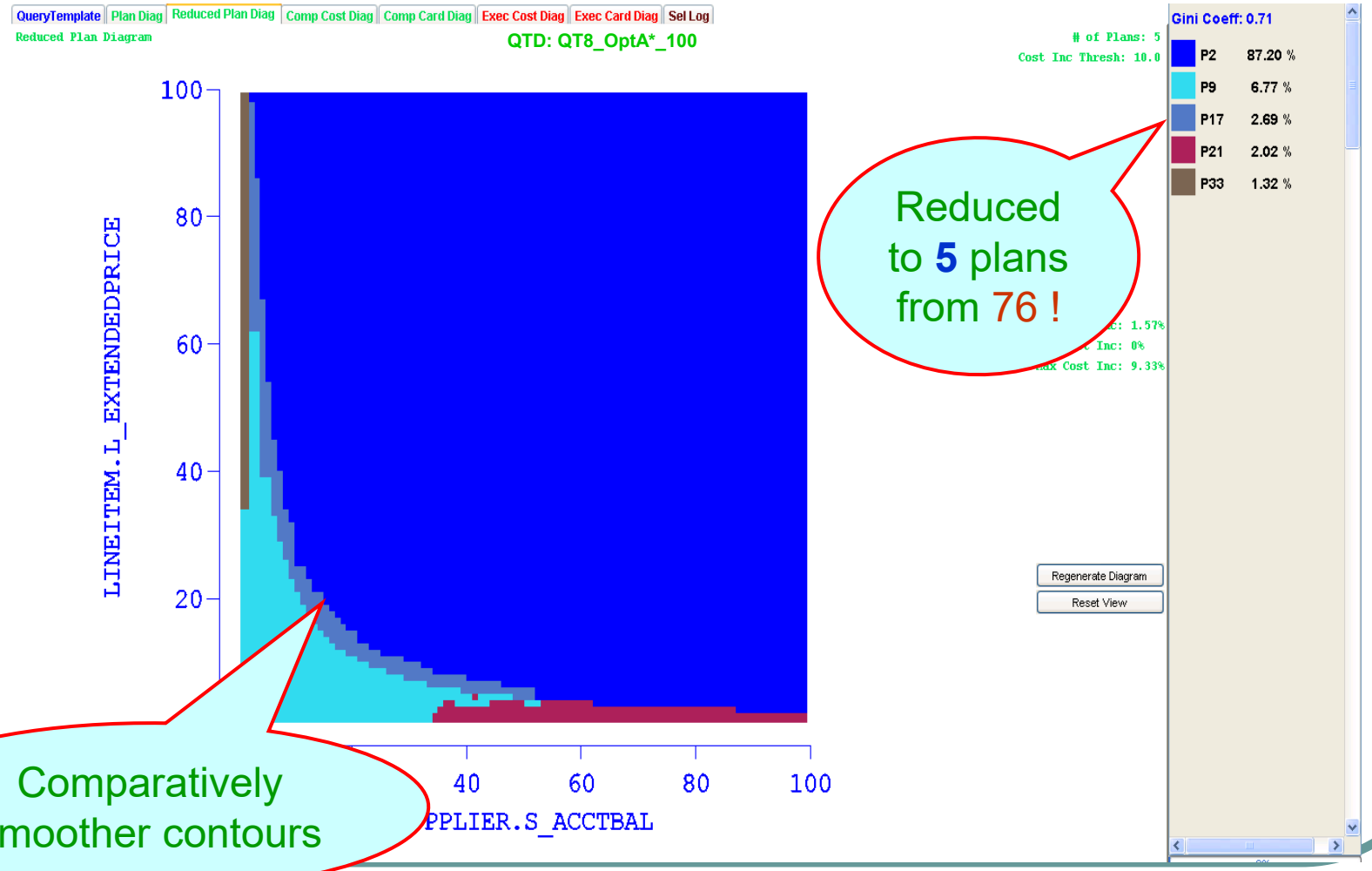
No query point in the original diagram has its estimated cost increased, post-swallowing, by more than λ percent (user-defined)

Analogy:

Sri Lanka agrees to be annexed by India if it is assured that the cost of living of each Lankan citizen is not increased by more than λ percent

Reduced Plan Diagram [λ=10%]

[QT8, OptA*, Res=100]



Is 10% increase acceptable?

- A 10% threshold is *well within* the confidence intervals of cost estimates of modern optimizers
- The degradation threshold is an *upper limit* – actual degradation is much lower in practice
- Traditional view is that a plan that is within *twice* of the optimal (i.e. λ = 100%) is "good"

PROBLEM ANALYSIS

Definition

- Plan diagram **P**
  m query points $q_1 \ldots q_m$
  n optimal plans $P_1 \ldots P_n$
- Each query point q_i
  - Selectivity location (*x%, y%*)
  - Cost of plan P_j at q_i is $c(P_j,q_i)$
  - Optimal plan $P_k \Rightarrow$ Color L_k
- Cost-increase threshold λ% (user defined)
- Reduced plan-diagram **R**:
  $L^R \subseteq L^P$

Problem: Find an **R** such that the number of plans (colors) in **R** is minimum subject to

$\forall\ P_k \in \mathbf{P}$, either

(a) $P_k \in \mathbf{R}$ *or*

(b) $\forall\ q \in P_k$, the assigned replacement plan $P_j \in \mathbf{R}$ is

s.t. $$\frac{c(P_j,q)}{c(P_k,q)} \leq 1 + \frac{\lambda}{100}$$

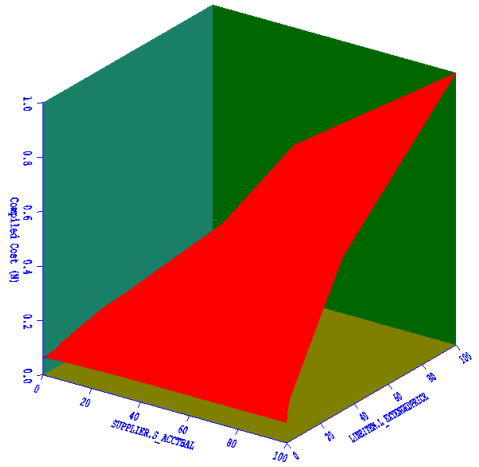
e.g. if λ = 10%, $$\frac{c(P_j,q)}{c(P_k,q)} \leq 1.1$$

Basic Requirement

- Need to be able to cost a plan P_k at points outside its own optimality region
  - called “Foreign Plan Costing” (FPC) or “Abstract Plan Costing”
- Option 1:
  - some optimizers natively support FPC feature
  - incurs non-trivial computational overheads
- Option 2:
  - use a conservative **cost-upper-bounding** approach
  - orders of magnitude faster

Option 2 Assumption: Plan Cost Monotonicity (PCM)

PCM: Cost distribution of each plan featured in plan diagram **P** is monotonically non-decreasing over entire selectivity space **S**.

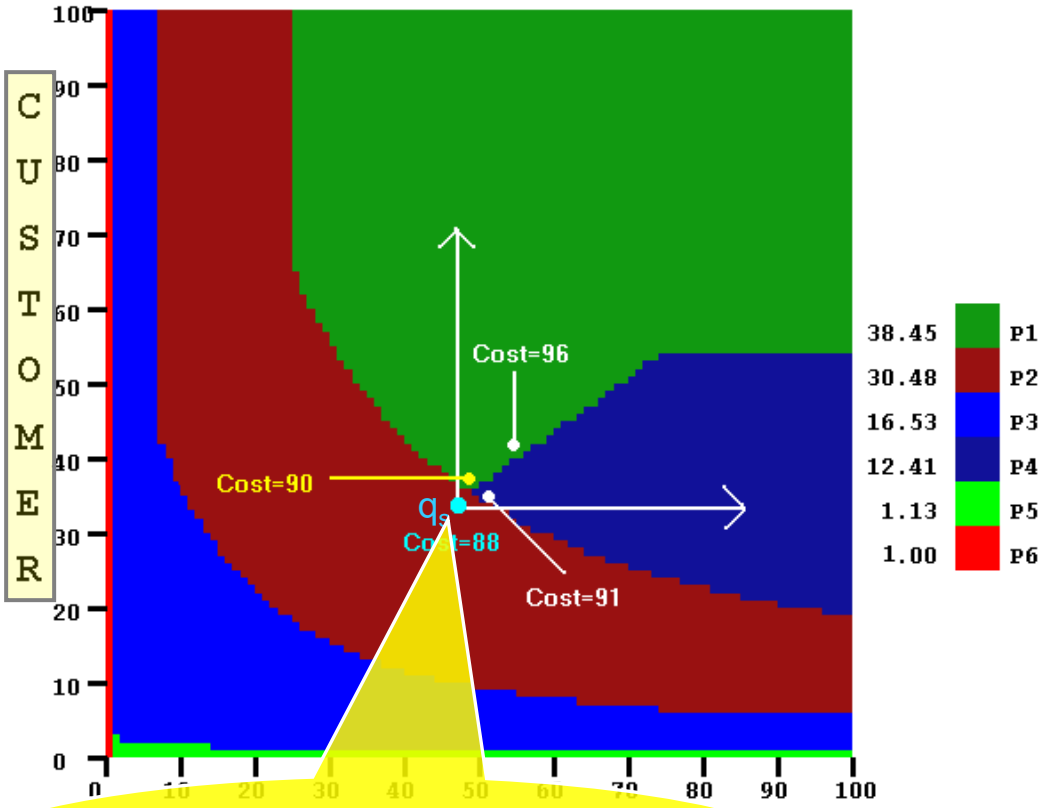


Cost function of plan P_k

True for most query templates since

selectivity↑ ⇒ input data↑ ⇒ query processing↑ ⇒ (est) cost↑

Cost-upper-bounding Approach

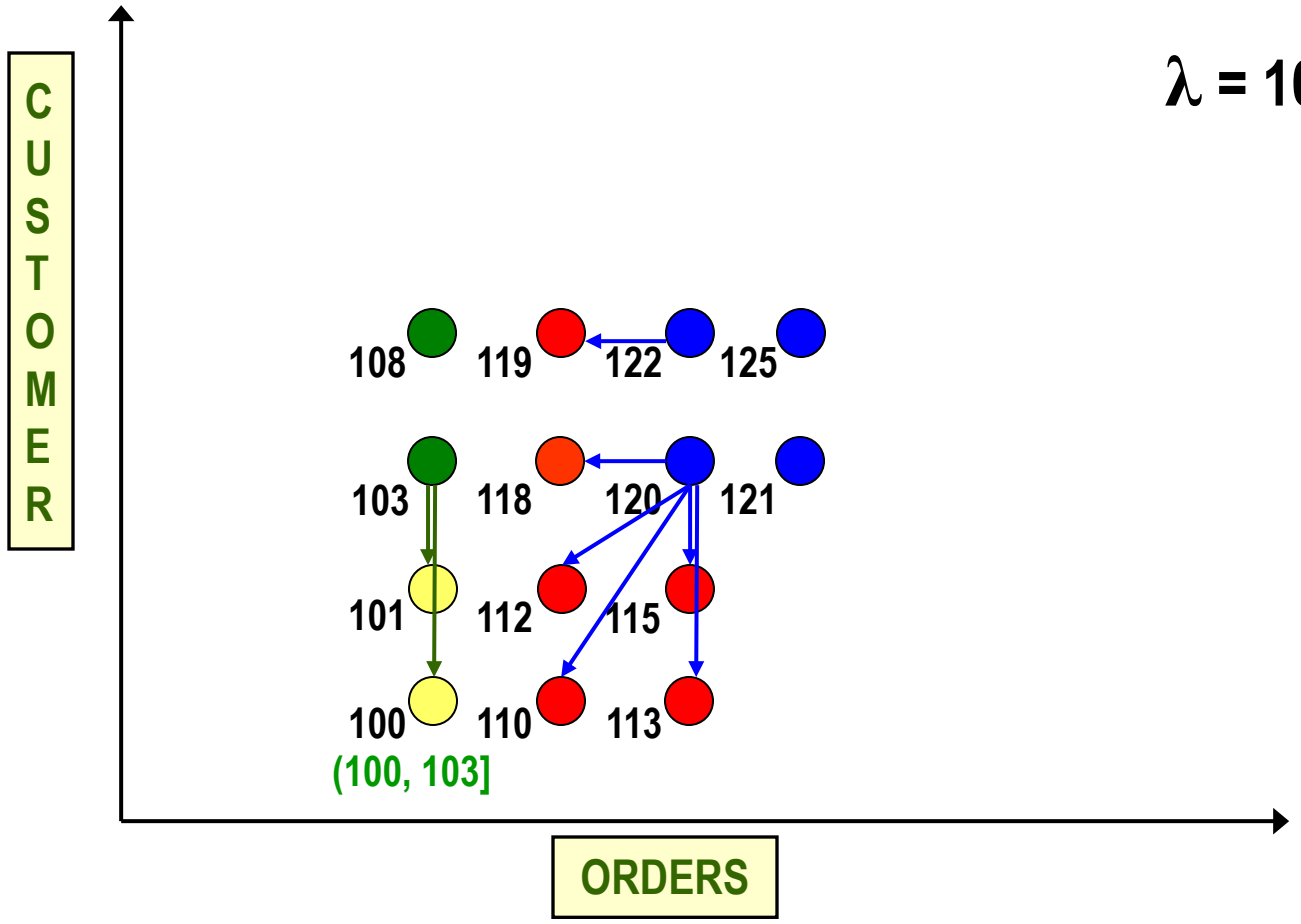


Cost of query point q_s with optimal plan P_2 is 88

PCM ⇒
Cost of a "foreign" query point in first quadrant of q_s is an upper bound on the cost of executing the foreign plan at q_s

⇒
Cost of executing q_s with foreign plans **P1** or **P4** lies in the intervals **(88, 90]** and **(88,91]**, respectively.

Example Plan Swallowing



Results

- Optimal plan diagram reduction (w.r.t. minimizing the number of plans/colors) is NP-hard
  - through problem-reduction from classical Set Cover

- Designed CostGreedy, a greedy heuristic-based algorithm with following properties:

  [m is number of query points, n is number of plans in diagram]

  - Time complexity is O(mn)
    - linear in number of plans for a given diagram resolution
  - Approximation Factor is O(ln m)
    - bound is both tight and optimal
    - in practice, closely approximates optimal

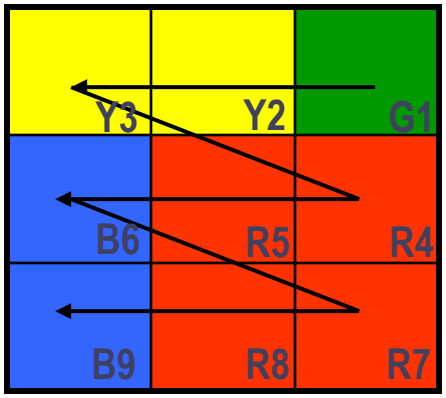
Cost Greedy Algorithm

- Assign a bin to each individual plan in **P**
- Start at the top right corner and proceed in reverse row-major order
  - first-quadrant info available when processing a query point
- Put a copy of each query point into all plan-bins (subsets) that it can belong to w.r.t. λ constraint: SetCover problem
- Iterative Greedy Criterion:
  - include in solution the plan (subset) that covers the maximum number of uncovered points
  - remove its covered points from all subsets and repeat until no uncovered points remain

Toy Example

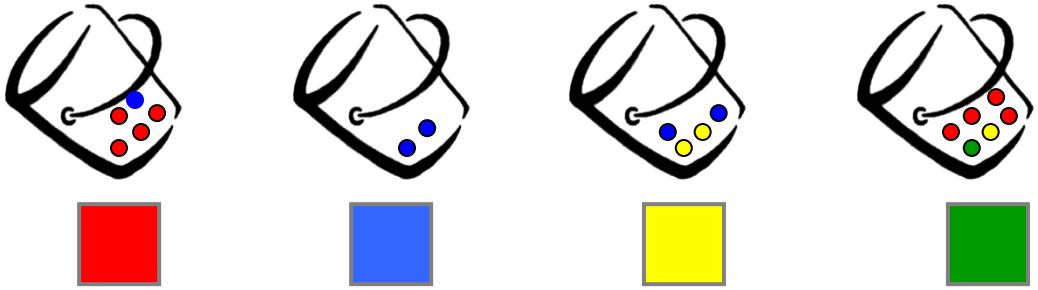
Plans in **R**

P



Pick this plan

Covers max (3) points

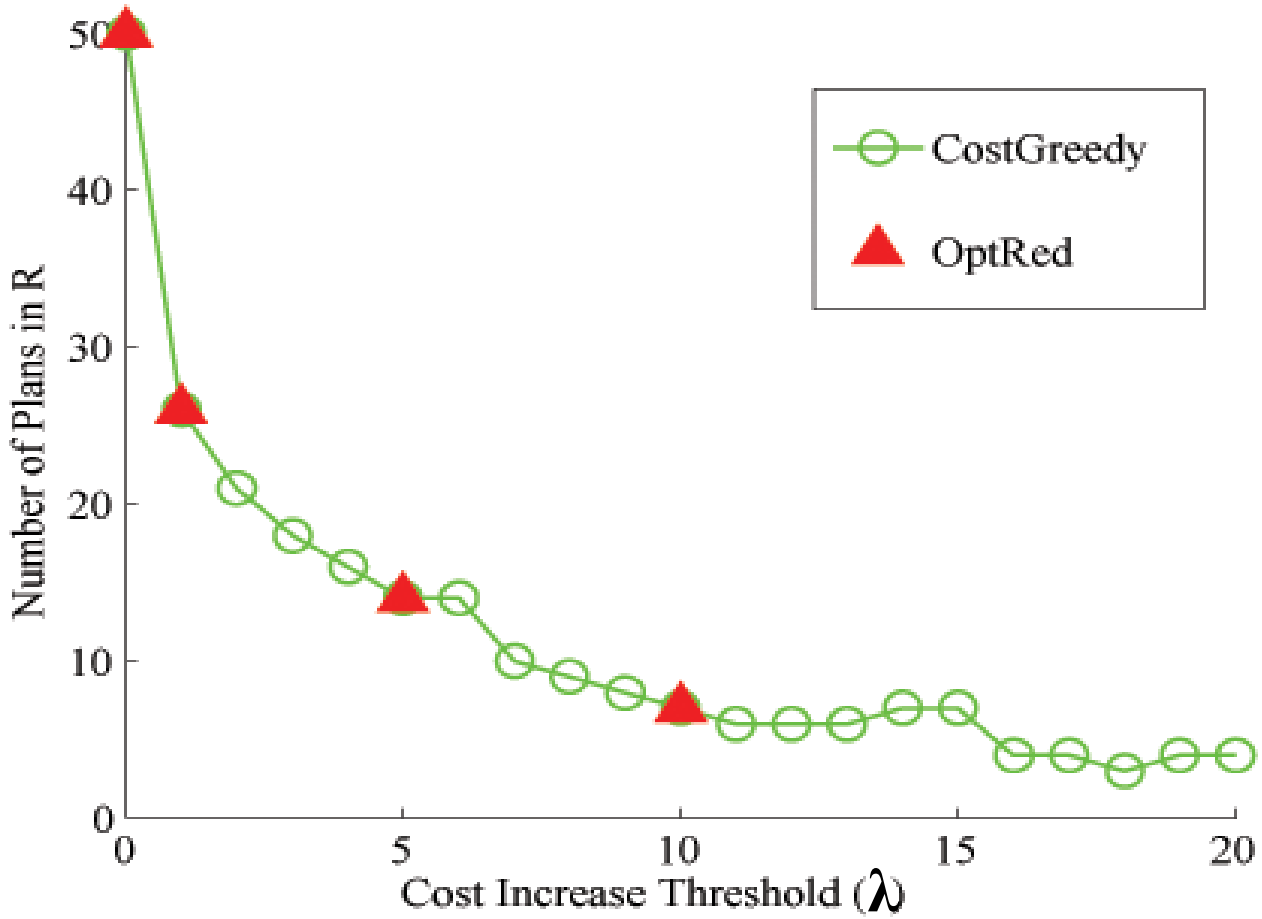


Computational Efficiency
[QT8, OptC]

Reduction Algorithm	Original # plans [100*100]	Reduced # plans (λ = 10%)	Time Taken	Original # plans [300*300]	Reduced # plans (λ = 10%)	Time Taken
Optimal Reduction	50	7	4 hours	89	*	* (time in years!)
CostGreedy	50	7	0.1 sec	89	6	3.2 sec

Reduction Quality

[QT8, OptC, Res = 100]



Anorexic Reduction

Extensive empirical evaluation with a spectrum of multi-dimensional TPCH-based query templates indicates that

"With a cost-increase-threshold of just 20%, virtually all complex plan diagrams
[irrespective of query templates, data distribution, query distribution, system configurations, etc.]

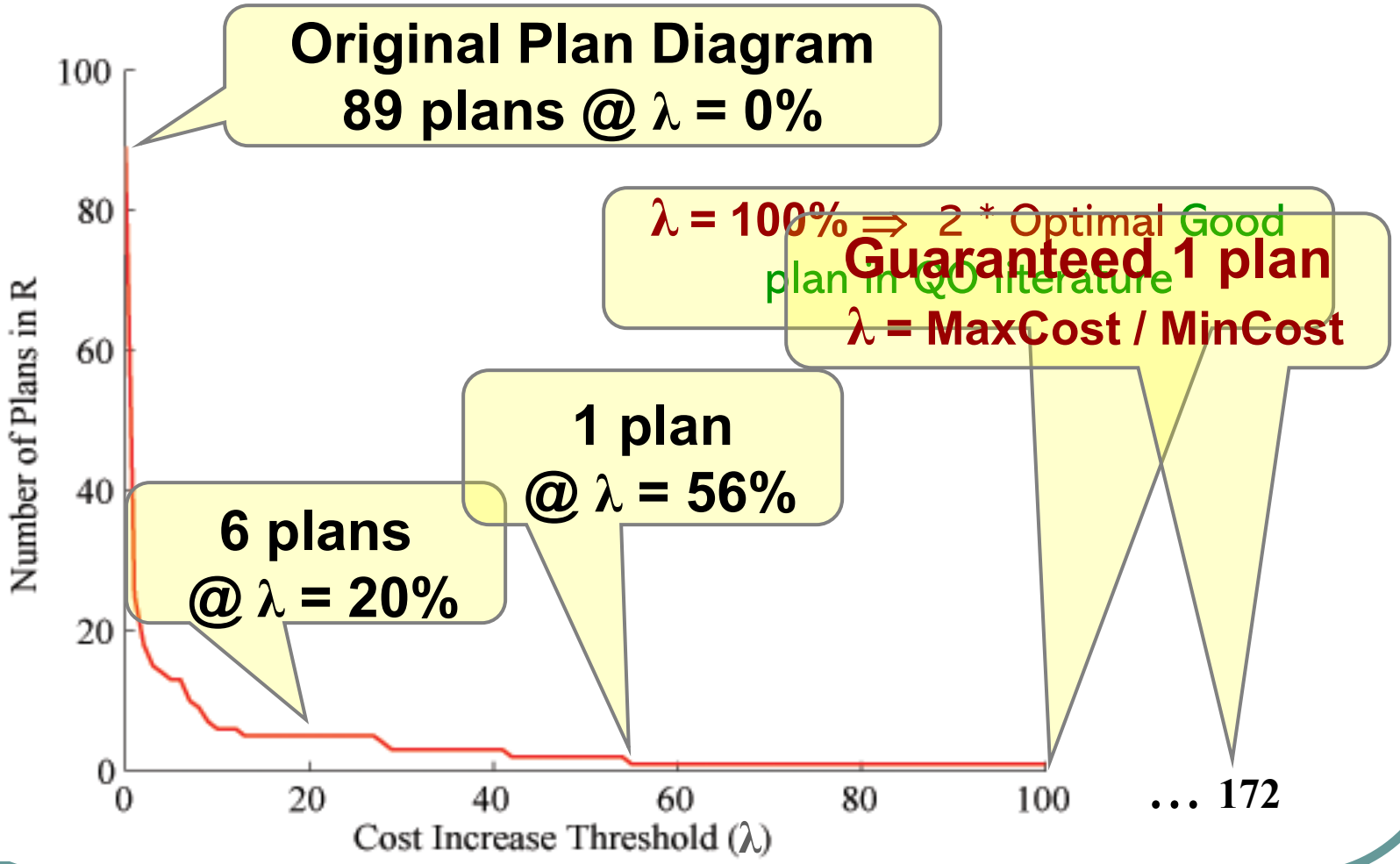
reduce to "anorexic levels" (~10 or less plans)!

Sample TPCH-based Results

[OptC, Res = 300]

TPC-H Query Template	Original # of Plans	Reduced Plans (λ = 10%)	Reduced Plans (λ = 20%)
2	76	20	12
5	31	10	6
8	89	6	6
9	91	9	4
10	31	6	4

Typical Graph of Plan Cardinality of R vs Cost Increase Threshold λ

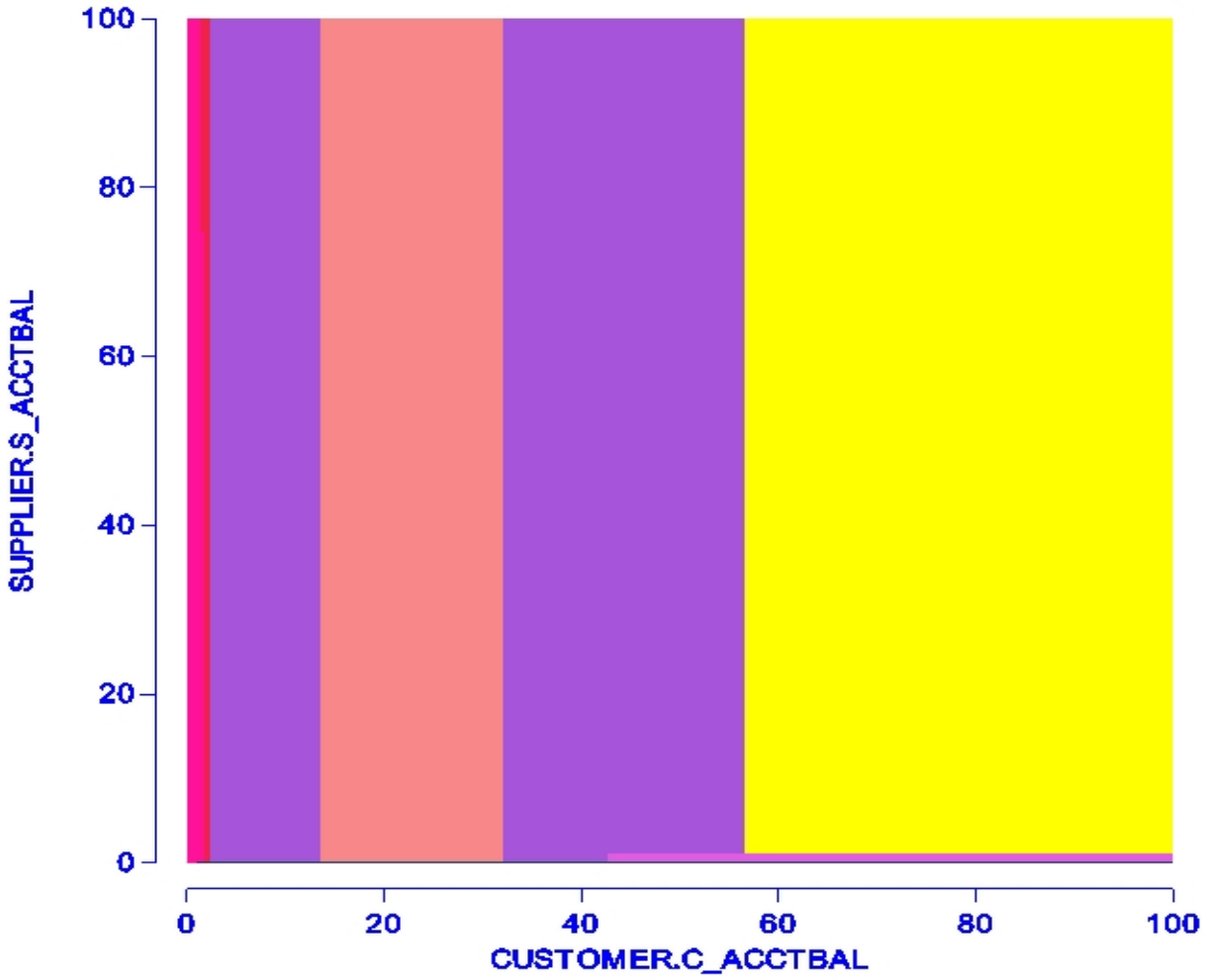


Reduction with Explicit Costing

- Results shown so far were conservative because of upper-bounding strategy
- With explicit swallower-plan costing, number of plans in reduced diagrams usually comes down to just a couple at 20% threshold!

Explicit Costing-based Reduction

[QT5, OptC, Res=30E, λ=20%]



Original Plan Diagram:
51 plans

Upper-bound Reduction:
7 plans

Explicit-cost Reduction:
3 plans

Problem Variant: Storage-Budgeted Plan Diagram Reduction

- Dual of plan-diagram reduction problem
- Problem Statement:
  Given **P** and storage constraint of retaining **≤ k** plans, choose the **k** plans so as to minimize the maximum cost-increase of swallowed query points in **R**.

- Optimal solution is NP-Hard
  - Karp reduction to Plan Diagram Reduction
- Threshold Greedy algorithm
  - Guaranteed to provide 2/3 of optimal benefit

Applications of Plan Diagram Reduction

- Quantifies redundancy in plan search space
- Provides better candidates for plan-cacheing
- Enhances viability of Parametric Query Optimization (PQO) techniques
- Improves efficiency/quality of Least-Expected-Cost (LEC) plans
- Minimizes overheads of multi-plan (e.g. Adaptive Query Processing) approaches
- Identifies selectivity-error resistant plan choices
  - retained plans are robust choices over larger selectivity parameter space

Parametric Query Optimization

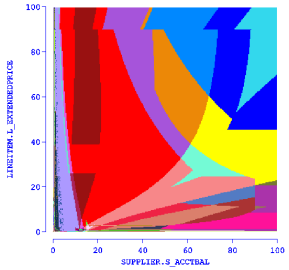
- Active research area for last two decades
  - VLDB 1992, 1998, 2002, 2003
  - IIT Kanpur (Sumit Ganguly & Co), IIT Bombay (Hulgeri & Sudarshan)
- Offline precompute, using geometric inferencing techniques, the parametric optimal set of plans (POSP) for the entire relational selectivity space
- At run-time, use actual selectivity values to identify the appropriate plan choice

Application to PQO

- Even if PQO assumptions true, need to store geometries of the plan diagram regions $\Rightarrow$ require spatial storage, R-tree index, etc.
- Cute alternative proposed by Hulgeri/Sudarshan: "*Cost* all *POSP plans at query location, and choose lowest cost plan*"
  - works only if number of plans in diagram is small, o/w overheads comparable to fresh optimization.
- Anorexic reduction ensures small cardinality in plan diagram

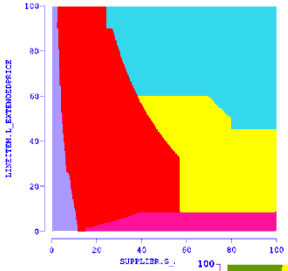
Take Away

PICASSO
[VLDB05]



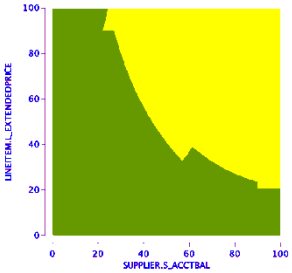
- Dense and Intricate Plan Diagrams
- PQO violations
- Optimizer bugs

Cost Greedy
[VLDB07]



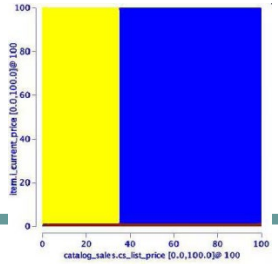
- Anorexic Reduction (less than10 plans)
- Local Near-optimality (20%)

SEER / GSPQO
[VLDB08]



- Anorexic Reduction
- Global Safety ("no harm")
- Robust Plans
- Efficient Approximation

EXPAND
[VLDB10]



- Anorexic Reduction
- Global Safety
- Robust Plans
- Online Processing

END PLAN DIAGRAM REDUCTION