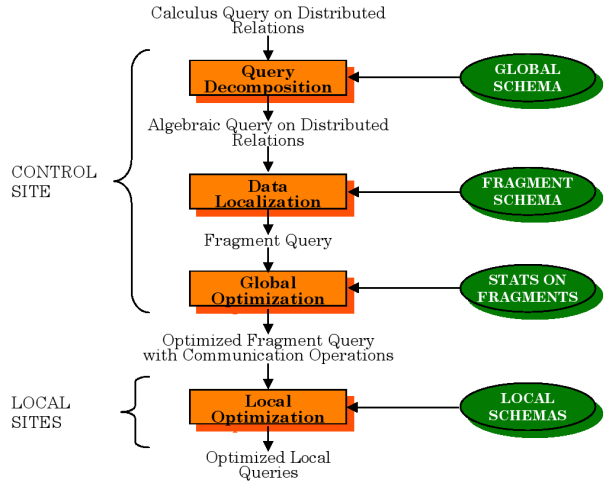
Distributed Database Systems
Fall 2018

Distributed Query Optimization
SL04

- Basic Concepts and Overview
- Distributed Cost Model
- Database Statistics
- Joins and Semijoins
- Query Optimization Algorithms

Basic Concepts/1

- **Query optimization:** Process of producing a (close to) optimal query execution plan that represents an execution strategy
  - The main task in query optimization is to consider different orderings of the operations

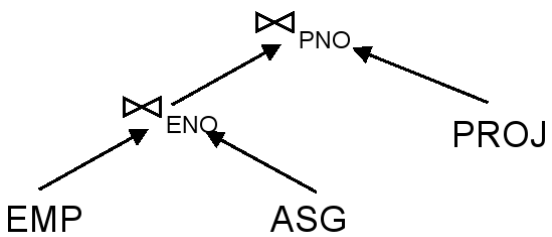


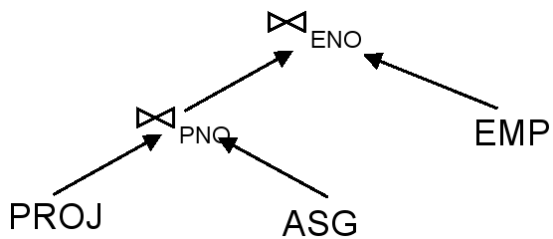
- Centralized query optimization:
  - Find (best) query execution plan in the space of equivalent query trees
  - Minimize an objective cost function
  - Gather statistics about relations
- Distributed query optimization brings additional issues
  - Linear query trees are not necessarily a good choice
  - Bushy query trees are not necessarily a bad choice
  - What and where to ship the relations
  - How to ship relations (ship as a whole, ship as needed)
  - When to use semi-joins instead of joins

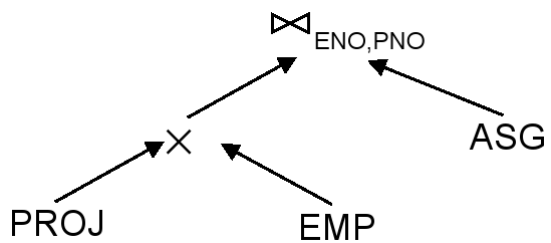
Basic Concepts/2

- **Search space:** The set of alternative query execution plans (query trees)
  - Typically very large
  - The main issue is to optimize joins
  - For N relations, there are $O(N!)$ equivalent join trees that can be obtained by applying commutativity and associativity rules
- **Example**: 3 equivalent query trees (join trees) of the joins in the following query

```
SELECT ENAME, RESP
FROM EMP, ASG, PROJ
WHERE EMP.ENO = ASG.ENO AND ASG.PNO = PROJ.PNO
```

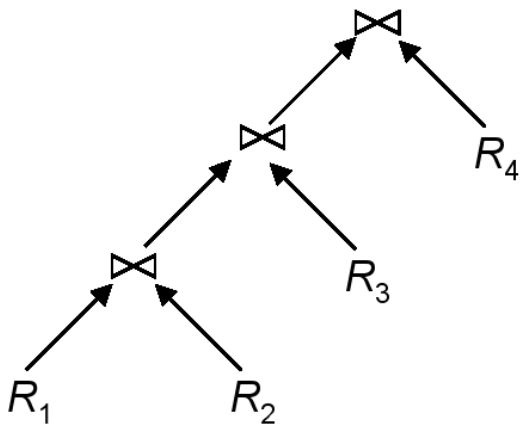




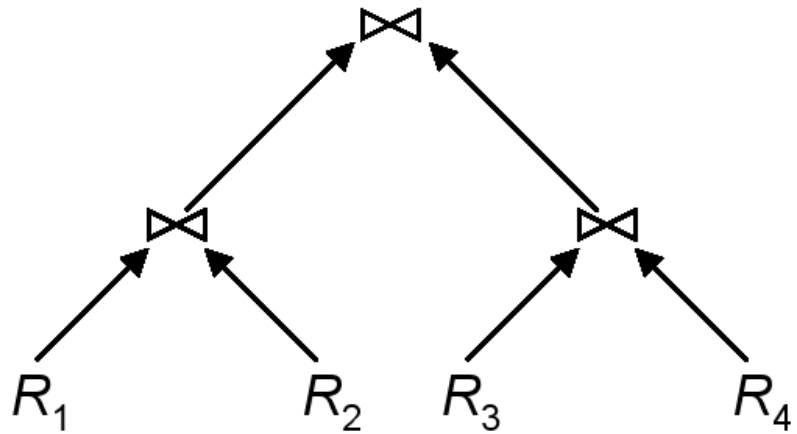


Basic Concepts/3

- **Reduction** of the search space
  - Restrict by means of heuristics
    - Perform unary operations before binary operations, etc
  - Restrict the shape of the join tree
    - Consider the type of trees (linear trees vs. bushy trees)



Linear Join Tree

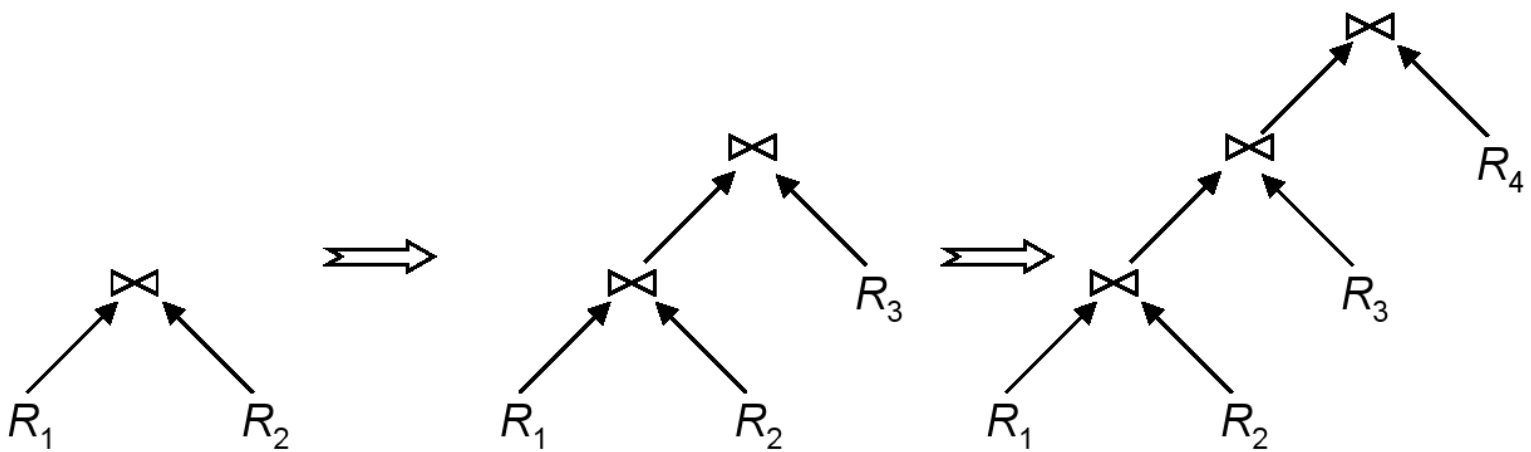


Bushy Join Tree

Basic Concepts/4

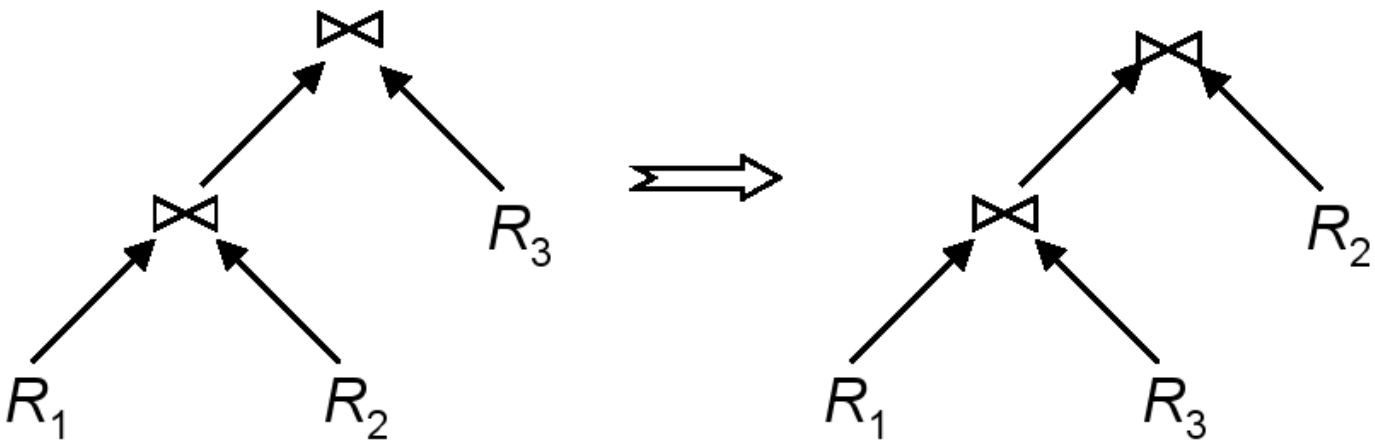
- There are two main strategies to **scan the search space**
  - Deterministic
  - Randomized

- **Deterministic scan** of the search space
  - Start from base relations and build plans by adding one relation at each step
  - Breadth-first strategy (BFS): build all possible plans before choosing the "best" plan (dynamic programming approach)
  - Depth-first strategy (DFS): build only one plan (greedy approach)



Basic Concepts/5

- **Randomized scan** of the search space
  - Search for optimal solutions around a particular starting point
  - e.g., iterative improvement or simulated annealing techniques
  - Trades optimization time for execution time
    - Does not guarantee that the best solution is obtained, but avoid the high cost of optimization
  - The strategy is better when more than 5-6 relations are involved



Basic Concepts/6

- **Ordering of the operators** of relational algebra is crucial for efficient query processing
- Rule of thumb: move expensive operators at the end of query processing
- Cost of RA operations:

Operation	**Complexity**
Select, Project (without duplicate elimination)	$O(n)$
Project (with duplicate elimination) Group	$O(n \log n)$
Join Semi-join Division Set Operators	$O(n \log n)$
Cartesian Product	$O(n^2)$

Query Optimization Overview/1

- **Query optimization** is a crucial and difficult part of the overall query processing
- Objective of query optimization is to **minimize** the total cost incurred at all sites:

$$\text{I/O cost} + \text{CPU cost} + \text{communication cost}$$

- Several issues have to be considered in query optimization
  - Types of query optimizers
    - wrt the search techniques (exhaustive search, heuristics)
    - wrt the time when the query is optimized (static, dynamic)
  - Statistics
  - Decision sites
  - Network topology
  - Use of semijoins

Query Optimization Overview/2

- **Types of Query Optimizers wrt Search Techniques**
  - Exhaustive search
    - Optimal
    - Cost-based
    - Combinatorial complexity in the number of relations
  - Heuristics
    - Not optimal
    - Regroups common sub-expressions
    - Performs selection, projection first
    - Extends a join with a series of semijoins
    - Reorders operations to reduce intermediate relation size
    - Optimizes individual operations

Query Optimization Overview/3

- **Types of Query Optimizers wrt Optimization Timing**
  - Static
    - Query is optimized prior to the execution
    - As a consequence it is difficult to estimate the size of the intermediate results
    - Typically amortizes over many executions
  - Dynamic
    - Optimization is done at runtime
    - Provides exact information on the intermediate relation sizes
    - Have to re-optimize for multiple executions
  - Hybrid
    - First, the query is compiled using a static algorithm
    - Then, if the error in estimate sizes greater than threshold, the query is re-optimized at run time

Query Optimization Overview/4

- **Statistics**
  - Relation/fragments
    - Cardinality
    - Size of a tuple
    - Fraction of tuples participating in a join with another relation/fragment
  - Attribute
    - Cardinality of domain
    - Actual number of distinct values
    - Distribution of attribute values (e.g., histograms)
  - Common assumptions
    - Independence between different attribute values
    - Uniform distribution of attribute values within their domain

Query Optimization Overview/5

- **Decision sites**
  - Centralized
    - Single site determines the "best" schedule
    - Simple
    - Knowledge about the entire distributed database is needed
  - Distributed
    - Cooperation among sites to determine the schedule
    - Only local information is needed
    - Cooperation comes with an overhead cost
  - Hybrid
    - One site determines the global schedule
    - Each site optimizes the local sub-queries

Query Optimization Overview/6

- **Network topology**
  - Wide area networks (WAN) - point-to-point
    - Characteristics: low bandwidth, low speed, high protocol overhead
    - Communication cost dominate; all other cost factors are ignored
    - Global schedule to minimize communication cost
    - Local schedules according to centralized query optimization
  - Local area networks (LAN)
    - Communication cost not that dominant
    - Total cost function should be considered
    - Broadcasting can be exploited (joins)
    - Special algorithms exist for star networks

Query Optimization Overview/9

- **Use of Semijoins**
  - Reduce the size of the join operands by first computing semijoins
  - Particularly relevant when the main cost is the communication cost
  - Improves the processing of distributed join operations by reducing the size of data exchange between sites
  - However, the number of messages as well as local processing time is increased

Distributed Cost Model/1

- Two different types of **cost functions** can be used
  - Reduce **total time**
    - Reduce each cost component (in terms of time) individually, i.e., do as little for each cost component as possible
    - Optimize the utilization of the resources (i.e., increase system throughput)
  - Reduce **response time**
    - Do as many things in parallel as possible
    - May increase total time because of increased total activity

Distributed Cost Model/2

- **Total time**: Sum of the time of all individual components
  - Local processing time: CPU time + I/O time
  - Communication time: fixed time to initiate a message + time to transmit the data

$$Total\_time = T_{CPU} * \#instructions + T_{I/O} * \#I/Os + T_{MSG} * \#messages + T_{TR} * \#bytes$$

- The individual components of the total cost have different weights:
  - Wide area network
    - Message initiation and transmission costs are high
    - Local processing cost is low (fast mainframes or minicomputers)
    - Ratio of communication to I/O costs is 20:1
  - Local area networks
    - Communication and local processing costs are more or less equal
    - Ratio of communication to I/O costs is 1:1.6 (10MB/s network)

Distributed Cost Model/3

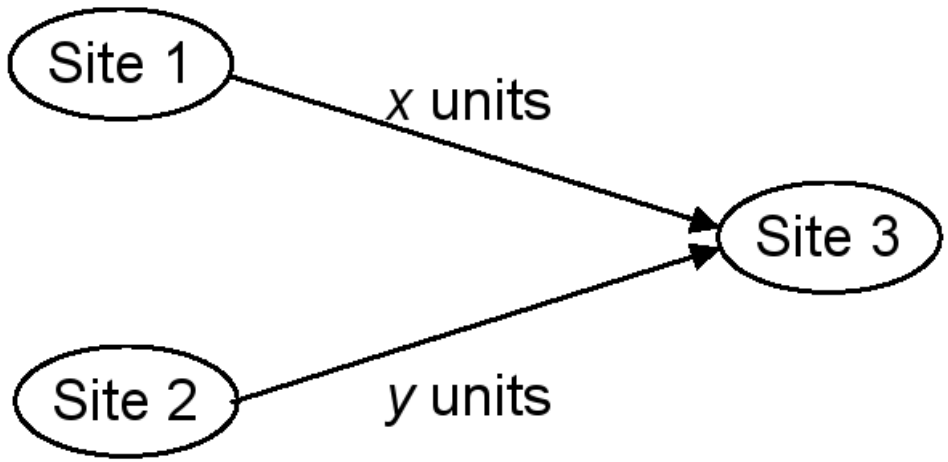
- **Response time**: Elapsed time between the initiation and the completion of a query

$$Response\_time = T_{CPU} * \#seq\_instructions + T_{I/O} * \#seq\_I/Os + T_{MSG} * \#seq\_messages + T_{TR} * \#seq\_bytes$$

  - where $\#seq\_x$ (x in instructions, I/O, messages, bytes) is the **maximum number** of x which must be done sequentially.
- Any processing and communication done in parallel is ignored

Distributed Cost Model/4

- **Example:** Query at site 3 with data from sites 1 and 2.



  - Assume that only the communication cost is considered
  - $Total\_time = T_{MSG} * 2 + T_{TR} * (x + y)$
  - $Response\_time = \max\{T_{MSG} + T_{TR} * x, T_{MSG} + T_{TR} * y\}$

Database Statistics/1

- The **primary cost factor** is the **size of intermediate relations**
  - that are produced during the execution and
  - must be transmitted over the network, if a subsequent operation is located on a different site
- It is costly to compute the size of the intermediate relations precisely.
- Instead **global statistics of relations and fragments** are computed and used to provide approximations

Database Statistics/2

- Let $R(A_1, A_2, \ldots, A_k)$ be a relation fragmented into $R_1, R_2, \ldots, R_r$.
- **Relation statistics**
  - min and max values of each attribute: $\min\{A_i\}$, $\max\{A_i\}$.
  - length of each attribute: $\mathit{length}(A_i)$
  - number of distinct values in each domain: $\mathit{card}(\mathit{dom}(A_i))$
- **Fragment statistics**
  - cardinality of the fragment: $\mathit{card}(R_i)$
  - cardinality of each attribute of each fragment: $\mathit{card}(\pi_{A_i}(R_j))$, $\mathit{card}(A_i)$

Database Statistics/3

- **Selectivity factor** of an operation: the proportion of tuples of an operand relation that participate in the result of that operation
- Assumption: independent attributes and uniform distribution of attribute values
- **Selectivity factor of a selection**

$$SF_\sigma(A = value) = \frac{1}{card(\pi_A(R))}$$

$$SF_\sigma(A > value) = \frac{\max(A) - value}{\max(A) - \min(A)}$$

$$SF_\sigma(A < value) = \frac{value - \min(A)}{\max(A) - \min(A)}$$

Database Statistics/4

- Properties of the selectivity factor of a selection

$$SF_\sigma(p(A_i) \wedge p(A_j)) = SF_\sigma(p(A_i)) * SF_\sigma(p(A_j))$$

$$SF_\sigma(p(A_i) \vee p(A_j)) = SF_\sigma(p(A_i)) + SF_\sigma(p(A_j)) - (SF_\sigma(p(A_i)) * SF_\sigma(p(A_j))$$

$$SF_\sigma(A \in \{values\}) = SF_\sigma(A = value) * card(\{values\})$$

Database Statistics/5

- **Cardinality** of intermediate results
  - Selection

$$card(\sigma_P(R)) = SF_\sigma(P) * card(R)$$

  - Projection
    - More difficult: correlations between projected attributes are unknown
    - Simple if the projected attribute is a key

$$card(\pi_A(R)) = card(R)$$

  - Cartesian Product

$$card(R \times S) = card(R) * card(S)$$

  - Union
    - upper bound: $card(R \cup S) \leq card(R) + card(S)$
    - lower bound: $card(R \cup S) \geq \max\{card(R), card(S)\}$
  - Set Difference
    - upper bound: $card(R - S) = card(R)$
    - lower bound: 0

Database Statistics/6

- **Selectivity factor** for joins

$$SF_{\bowtie} = \frac{card(R \bowtie S)}{card(R) * card(S)}$$

- **Cardinality** of joins
  - Upper bound: cardinality of Cartesian Product

$$card(R \bowtie S) \leq card(R) * card(S)$$

  - General case (if SF is given):

$$card(R \bowtie S) = SF_{\bowtie} * card(R) * card(S)$$

  - Special case: $R.A$ is a key of R and $S.A$ is a foreign key of S;
    - each S-tuple matches with at most one tuple of R

$$card(R \bowtie_{R.A=S.A} S) = card(S)$$

Database Statistics/7

- **Selectivity factor** for semijoins: fraction of R-tuples that join with S-tuples
  - An approximation is the selectivity of A in S

$$SF_{\ltimes}(R \ltimes_A S) = SF_{\ltimes}(S.A) = \frac{card(\pi_A(S))}{card(dom[A])}$$

- **Cardinality** of semijoin (general case):

$$card(R \ltimes_A S) = SF_{\ltimes}(S.A) * card(R)$$

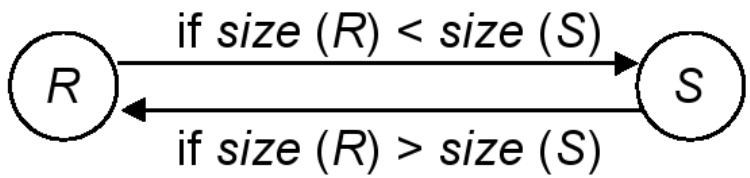
- Example: $R.A$ is a foreign key in S ($S.A$ is a primary key)
  Then $SF = 1$ and the result size corresponds to the size of R

Join Ordering in Fragment Queries/1

- **Join ordering** is an important aspect in centralized DBMS, and it **is even more important in a DDBMS** since joins between fragments that are stored at different sites may increase the communication time.
- Two approaches exist:
  - Optimize the ordering of joins directly
    - INGRES and distributed INGRES
    - System R and System R^*
  - Replace joins by combinations of semijoins in order to minimize the communication costs
    - Hill Climbing and SDD-1

Join Ordering in Fragment Queries/2

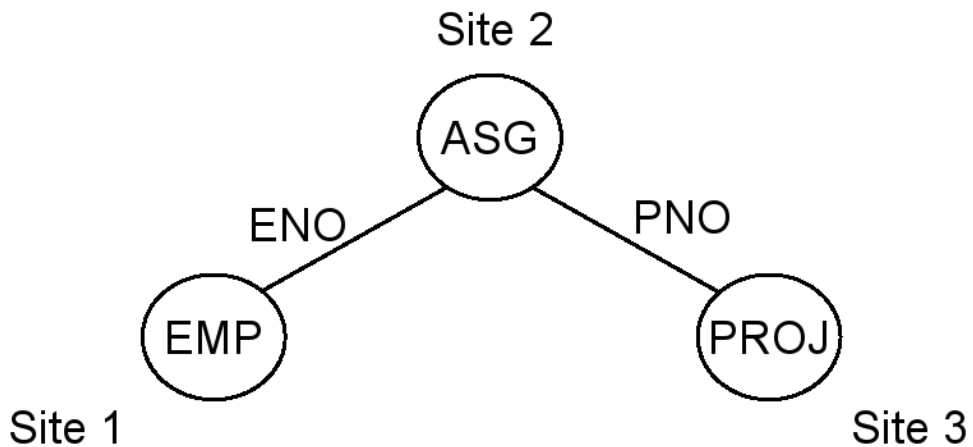
- **Direct join odering** of two relation/fragments located at different sites
  - Move the smaller relation to the other site
  - We have to estimate the size of R and S



Join Ordering in Fragment Queries/3

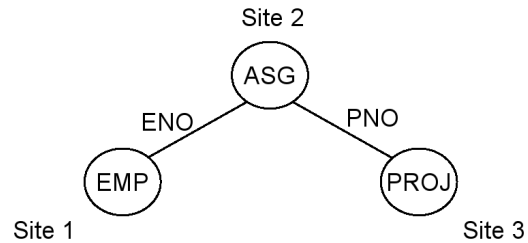
- **Direct join ordering** of queries involving more than two relations is substantially more complex
- **Example:** Consider the following query and the respective join graph, where we make also assumptions about the locations of the three relations/fragments

$$PROJ \bowtie_{PNO} ASG \bowtie_{ENO} EMP$$



Join Ordering in Fragment Queries/4

- **Example (contd.):** The query can be evaluated in at least 5 different ways.
  - Plan 1:
    EMP→Site 2
    Site 2: EMP'=EMP⋈ASG
    EMP'→Site 3
    Site 3: EMP'⋈PROJ
  - Plan 2:
    ASG→Site 1
    Site 1: EMP'=EMP⋈ASG
    EMP'→Site 3
    Site 3: EMP'⋈PROJ
  - Plan 3:
    ASG→Site 3
    Site 3: ASG'=ASG⋈PROJ
    ASG'→Site 1
    Site 1: ASG'⋈EMP



  - Plan 4:
    PROJ→Site 2
    Site 2: PROJ'=PROJ⋈ASG
    PROJ'→Site 1
    Site 1: PROJ'⋈EMP
  - Plan 5:
    EMP→Site 2
    PROJ→Site 2
    Site 2: EMP⋈PROJ⋈ASG

Join Ordering in Fragment Queries/5

- To select a plan, a lot of information is needed, including
  - $size(EMP)$
  - $size(ASG)$
  - $size(PROJ)$
  - $size(EMP \bowtie ASG)$
  - $size(ASG \bowtie PROJ)$
  - Possibilities of parallel execution if response time is used

Semijoin Based Algorithms/1

- **Semijoins** can be used to efficiently implement joins
  - The semijoin acts as a size reducer (similar as to a selection) such that smaller relations need to be transferred
- Consider two relations: R located at site 1 and S located and site 2
  - Solution with semijoins: Replace one or both operand relations/fragments by a semijoin, using the following rules:

$$
\begin{aligned}
R \bowtie_A S &\iff (R \ltimes_A S) \bowtie_A S \\
&\iff R \bowtie_A (S \ltimes_A R) \\
&\iff (R \ltimes_A S) \bowtie_A (S \ltimes_A R)
\end{aligned}
$$

- The semijoin is beneficial if the cost to produce and send the required data to the other site is less than the cost of sending the whole operand relation and doing the actual join.

Semijoin Based Algorithms/2

- **Cost analysis** $R \bowtie_A S$ vs. $(R \ltimes_A S) \bowtie S$, assuming that $size(R) < size(S)$
  - Perform the join $R \bowtie S$:
    - $R \rightarrow$ Site 2
    - Site 2 computes $R \bowtie S$
  - Perform the semijoins $(R \ltimes S) \bowtie S$:
    - $S' = \pi_A(S)$
    - $S' \rightarrow$ Site 1
    - Site 1 computes $R' = R \ltimes S'$
    - $R' \rightarrow$ Site 2
    - Site 2 computes $R' \bowtie S$
  - Semijoin is better if: $size(\pi_A(S)) + size(R \ltimes S') < size(R)$
- The **semijoin** approach is better if the semijoin acts as a **sufficient reducer** (i.e., a few tuples of R participate in the join)
- The **join** approach is better if **almost all tuples of R participate** in the join

INGRES Algorithm/1

- **INGRES** uses a dynamic query optimization algorithm that recursively breaks a query into smaller pieces. It is based on the following ideas:
  - An n-relation query q is **decomposed** into n subqueries $q_1 \rightarrow q_2 \rightarrow \cdots \rightarrow q_n$
    - Each q_i is a mono-relation (mono-variable) query
    - The output of q_i is consumed by q_{i+1}
  - For the decomposition two basic techniques are used: **detachment** and **substitution**
  - There is a processor that can **efficiently** process mono-relation queries
    - Optimizes each query independently for the access to a single relation

INGRES Algorithm/2

- **Detachment:** Break a query q into $q' \rightarrow q''$, based on a common relation that is the result of q', i.e.
  - The query $q =$

    SELECT $R_2.A_2, \ldots, R_n.A_n$
    FROM $R_1, R_2, \ldots, R_n$
    WHERE $P_1(R_1.A'_1)$
    AND $P_2(R_1.A_1, \ldots, R_n.A_n)$

  - is decomposed by detachment of the common relation R_1 into $q' =$

    SELECT $R_1.A_1$ **INTO** R'_1
    FROM R_1
    WHERE $P_1(R_1.A'_1)$

  - and query $q'' =$

    SELECT $R_2.A_2, \ldots, R_n.A_n$
    FROM $R'_1, R_2, \ldots, R_n$
    WHERE $P_2(R'_1.A_1, \ldots, R_n.A_n)$

- Detachment **reduces the size** of the relation on which q'' is defined.

INGRES Algorithm/3

- **Example:** Consider query $q1$: "Names of employees working on the CAD/CAM project"

$q_1 =$
```
SELECT EMP.ENAME
FROM EMP, ASG, PROJ
WHERE EMP.ENO = ASG.ENO
AND ASG.PNO = PROJ.PNO
AND PROJ.PNAME = 'CAD/CAM'
```

- Decompose q_1 into $q_{11} \rightarrow q'$:

$q_{11} =$
```
SELECT PROJ.PNO INTO JVAR
FROM PROJ
WHERE PROJ.PNAME = 'CAD/CAM'
```

$q' =$
```
SELECT EMP.ENAME
FROM EMP, ASG, JVAR
WHERE EMP.ENO = ASG.ENO
AND ASG.PNO = JVAR.PNO
```

INGRES Algorithm/4

- **Example (contd.):** The successive detachments may transform q' into $q_{12} \rightarrow q_{13}$:

$q' =$
```
SELECT EMP.ENAME
FROM EMP, ASG, JVAR
WHERE EMP.ENO = ASG.ENO
AND ASG.PNO = JVAR.PNO
```

$q_{12} =$
```
SELECT ASG.ENO INTO GVAR
FROM ASG, JVAR
WHERE ASG.PNO = JVAR.PNO
```

$q_{13} =$
```
SELECT EMP.ENAME
FROM EMP, GVAR
WHERE EMP.ENO = GVAR.ENO
```

- q_1 is now decomposed by detachment into $q_{11} \rightarrow q_{12} \rightarrow q_{13}$
- q_{11} is a mono-relation query
- q_{12} and q_{13} are multi-relation queries, which cannot be further detached; also called **irreducible**

INGRES Algorithm/5

- **Tuple substitution** allows to convert an irreducible query q into mono-relation queries.
  - Choose a relation R_1 in q for tuple substitution
  - For each tuple in R_1, replace the R_1-attributes referred in q by their actual values, thereby generating a set of subqueries q' with $n-1$ relations, i.e.,

    $$q(R_1, R_2, \ldots, R_n) \text{ is replaced by } \{q'(t_{1_i}, R_2, \ldots, R_n), t_{1_i} \in R_1\}$$

- **Example (contd.):** Assume *GVAR* consists only of the tuples $\{E1, E2\}$. Then q_{13} is rewritten with tuple substitution in the following way:

$q_{13} =$
```
SELECT EMP.ENAME
FROM EMP, GVAR
WHERE EMP.ENO = GVAR.ENO
```

$q_{131} =$
```
SELECT EMP.ENAME
FROM EMP
WHERE EMP.ENO = 'E1'
```

INGRES Algorithm/6

- **Example (contd.):**

  q_{132} =
  ```
  SELECT EMP.ENAME
  FROM EMP
  WHERE EMP.ENO = 'E2'
  ```

  - q_{131} and q_{132} are mono-relation queries

Distributed INGRES Algorithm

- The **distributed INGRES query optimization algorithm** is very similar to the centralized INGRES algorithm.
  - In addition to the centralized INGRES, the distributed one should break up each query q_i into sub-queries that operate on fragments; only horizontal fragmentation is handled.
  - Optimization with respect to a combination of communication cost and response time

System R Algorithm/1

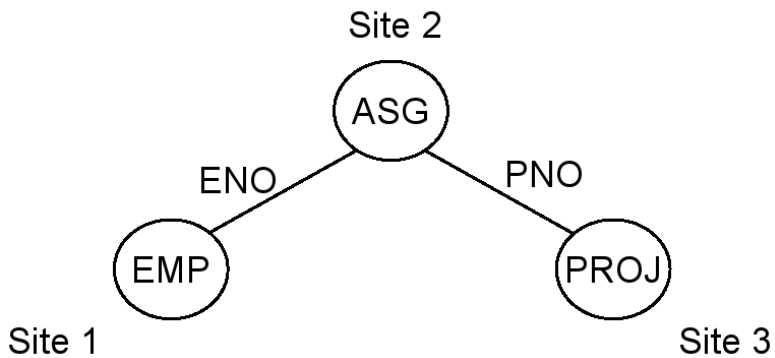
- The **System R** (centralized) query optimization algorithm
  - Performs static query optimization based on "exhaustive search" of the solution space and a cost function (IO cost + CPU cost)
    - Input: relational algebra tree
    - Output: optimal relational algebra tree
    - Dynamic programming technique is applied to reduce the number of alternative plans
  - The **optimization algorithm** consists of two steps
    1. Predict the best access method to each individual relation (mono-relation query)
    2. Consider using index, file scan, etc.
    3. For each relation R, estimate the best join ordering
    4. R is first accessed using its best single-relation access method
    5. Efficient access to inner relation is crucial
  - Considers two different join strategies
    - (Indexed-) nested loop join
    - Sort-merge join

System R Algorithm/2

- **Example:** Consider query $q1$: "*Names of employees working on the CAD/CAM project*"

$$PROJ \bowtie_{PNO} ASG \bowtie_{ENO} EMP$$

  - Join graph



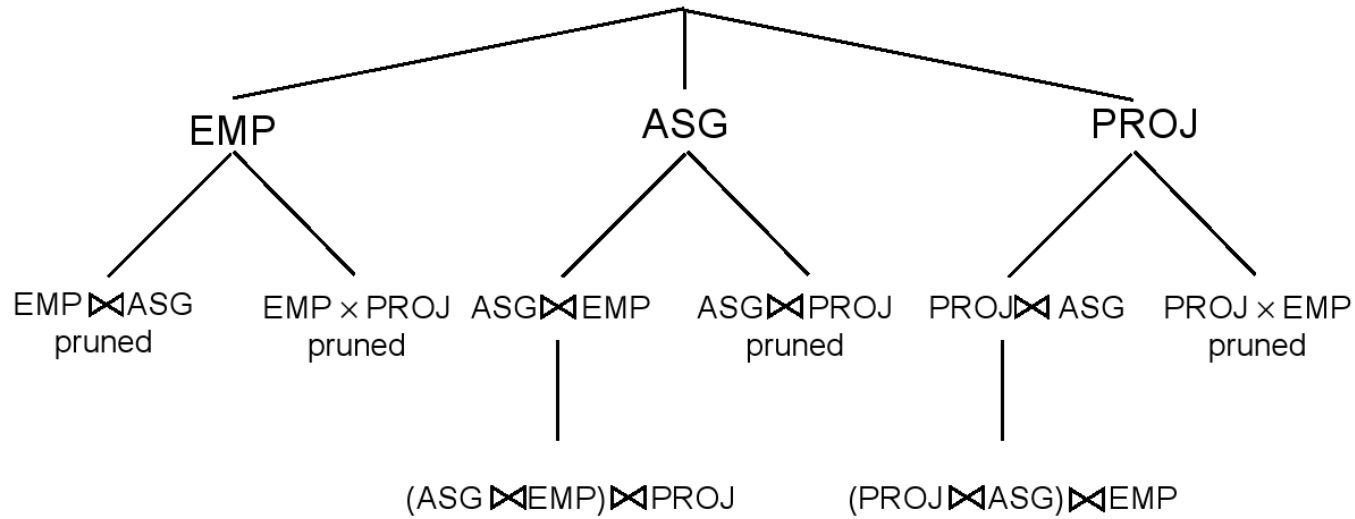
  - Indexes
    - EMP has an index on ENO
    - ASG has an index on PNO
    - PROJ has an index on PNO and an index on PNAME

System R Algorithm/3

- **Example (contd.):** Step 1 – Select the best single-relation access paths
  - EMP: sequential scan (because there is no selection on EMP)
  - ASG: sequential scan (because there is no selection on ASG)
  - PROJ: index on PNAME (because there is a selection on PROJ based on PNAME)

System R Algorithm/4

- **Example (contd.):** Step 2 – Select the best join ordering for each relation



  - (EMP × PROJ) and (PROJ × EMP) are pruned since they are CPs
  - (ASG ⋈ PROJ) pruned because (we assume) it has higher cost than (PROJ ⋈ ASG); similar for (ASG ⋈ EMP)
  - Best total join order ((PROJ⋈ ASG)⋈ EMP), since it uses the indexes best
    - Select PROJ using index on PNAME
    - Join with ASG using index on PNO
    - Join with EMP using index on ENO

Distributed System R^* Algorithm/1

- The **System R^* query optimization** algorithm is an extension of the System R query optimization algorithm with the following main characteristics:
  - Only the whole relations can be distributed, i.e., fragmentation and replication is not considered
  - Query compilation is a distributed task, coordinated by a **master site**, where the query is initiated
  - Master site makes all inter-site decisions, e.g., selection of the execution sites, join ordering, method of data transfer, ...
  - The **local sites** do the intra-site (local) optimizations, e.g., local joins, access paths
- Join ordering and data transfer between different sites are the most critical issues to be considered by the master site

Distributed System R^* Algorithm/2

- Two methods for **inter-site data transfer**
  - **Ship whole:** The entire relation is shipped to the join site and stored in a temporary relation
    - Larger data transfer
    - Smaller number of messages
    - Better if relations are small
  - **Fetch as needed:** The outer relation is sequentially scanned, and for each tuple the join value is sent to the site of the inner relation and the matching inner tuples are sent back (i.e., semijoin)
    - Number of messages = O(cardinality of outer relation)
    - Data transfer per message is minimal
    - Better if relations are large and the selectivity is good

Distributed System R^* Algorithm/3

- Four main **join strategies** for $R \bowtie S$:
  - R is outer relation
  - S is inner relation
- Notation:
  - LT denotes local processing time
  - CT denotes communication time
  - s denotes the average number of S-tuples that match an R-tuple

- **Strategy 1:** Ship the entire outer relation to the site of the inner relation, i.e.,
  - Retrieve outer tuples
  - Send them to the inner relation site
  - Join them as they arrive

$$\begin{aligned} Total\_cost = {} & LT(\text{retrieve } card(R) \text{ tuples from } R) + \\ & CT(size(R)) + \\ & LT(\text{retrieve } s \text{ tuples from } S) * card(R) \end{aligned}$$

Distributed System R^* Algorithm/4

- **Strategy 2:** Ship the entire inner relation to the site of the outer relation. We cannot join as they arrive; they need to be stored.
  - The inner relation S need to be stored in a temporary relation

$$\begin{aligned} Total\_cost = {} & LT(\text{retrieve } card(S) \text{ tuples from } S) + \\ & CT(size(S)) + \\ & LT(\text{store } card(S) \text{ tuples in } T) + \\ & LT(\text{retrieve } card(R) \text{ tuples from } R) + \\ & LT(\text{retrieve } s \text{ tuples from } T) * card(R) \end{aligned}$$

Distributed System R^* Algorithm/5

- **Strategy 3:** Fetch tuples of the inner relation as needed for each tuple of the outer relation.
  - For each R-tuple, the join attribute A is sent to the site of S
  - The s matching S-tuples are retrieved and sent to the site of R

$$\begin{aligned} Total\_cost = {} & LT(\text{retrieve } card(R) \text{ tuples from } R) + \\ & CT(length(A)) * card(R) + \\ & LT(\text{retrieve } s \text{ tuples from } S) * card(R) + \\ & CT(s * length(S)) * card(R) \end{aligned}$$

Distributed System R^* Algorithm/6

- **Strategy 4:** Move both relations to a third site and compute the join there.
  - The inner relation S is first moved to a third site and stored in a temporary relation.
  - Then the outer relation is moved to the third site and its tuples are joined as they arrive.

$$\begin{aligned} Total\_cost = {} & LT(\text{retrieve } card(S) \text{ tuples from } S) + \\ & CT(size(S)) + \\ & LT(\text{store } card(S) \text{ tuples in } T) + \\ & LT(\text{retrieve } card(R) \text{ tuples from } R) + \\ & CT(size(R)) + \\ & LT(\text{retrieve } s \text{ tuples from } T) * card(R) \end{aligned}$$

Hill-Climbing Algorithm/1

- **Hill-Climbing query optimization** algorithm
  - Refinements of an initial feasible solution are recursively computed until no more cost improvements can be made
  - Semijoins, data replication, and fragmentation are not used
  - Devised for wide area point-to-point networks
  - The first distributed query processing algorithm

Hill-Climbing Algorithm/2

- The hill-climbing algorithm proceeds as follows
  1. Select initial feasible execution strategy ES0
     - i.e., a global execution schedule that includes all intersite communication
     - Determine the candidate result sites, where a relation referenced in the query exist
     - Compute the cost of transferring all the other referenced relations to each candidate site
     - ES0 = candidate site with minimum cost
  2. Split ES0 into two strategies: ES1 followed by ES2
     - ES1: send one of the relations involved in the join to the other relation's site
     - ES2: send the join result to the final result site
  3. Replace ES0 with the split schedule which gives

$$cost(ES1) + cost(\text{local join}) + cost(ES2) < cost(ES0)$$

  4. Recursively apply steps 2 and 3 on ES1 and ES2 until no more benefit can be gained
  5. Check for redundant transmissions in the final plan and eliminate them

Hill-Climbing Algorithm/3

- **Example:** *What are the salaries of engineers who work on the CAD/CAM project?*

$$\pi_{SAL}(PAY \bowtie_{TITLE} EMP \bowtie_{ENO} (ASG \bowtie_{PNO} (\sigma_{PNAME="CAD/CAM"}(PROJ))))$$

  - Schemas: EMP(ENO, ENAME, TITLE), ASG(ENO, PNO, RESP, DUR), PROJ(PNO, PNAME, BUDGET, LOC), PAY(TITLE, SAL)
  - Statistics

Relation	Size	Site
EMP	8	1
PAY	4	2
PROJ	1	3
ASG	10	4

  - Assumptions:
    - Size of relations is defined as their cardinality
    - Minimize total cost
    - Transmission cost between two sites is 1
    - Ignore local processing cost
    - size(EMP $\bowtie$ PAY) = 8, size(PROJ $\bowtie$ ASG) = 2, size(ASG $\bowtie$ EMP) = 10

Hill-Climbing Algorithm/4

- **Example (contd.):** Determine initial feasible execution strategy
  - Alternative 1: Resulting site is site 1

$$\begin{aligned} Total\_cost &= cost(\text{PAY} \rightarrow \text{Site1}) + cost(\text{ASG} \rightarrow \text{Site1}) + \\ &\quad cost(\text{PROJ} \rightarrow \text{Site1}) \\ &= 4 + 10 + 1 = 15 \end{aligned}$$

  - Alternative 2: Resulting site is site 2

$$\text{Total cost} = 8 + 10 + 1 = 19$$

  - Alternative 3: Resulting site is site 3

$$\text{Total cost} = 8 + 4 + 10 = 22$$

  - Alternative 4: Resulting site is site 4

$$\text{Total cost} = 8 + 4 + 1 = 13$$

  - Therefore ES0 = EMP$\rightarrow$Site4; PAY $\rightarrow$ Site4; PROJ $\rightarrow$ Site4

Hill-Climbing Algorithm/5

- **Example (contd.):** Candidate split
  - Alternative 1: ES1, ES2, ES3
    - ES1: EMP→Site 2
    - ES2: (EMP⋈PAY) → Site4
    - ES3: PROJ→Site 4

$$\begin{aligned} Tot\_cost &= cost(\text{EMP} \rightarrow \text{Site2}) + \\ &\quad cost((\text{EMP} \bowtie \text{PAY}) \rightarrow \text{Site4}) + \\ &\quad cost(\text{PROJ} \rightarrow \text{Site4}) \\ &= 8 + 8 + 1 = 17 \end{aligned}$$

  - Alternative 2: ES1, ES2, ES3
    - ES1: PAY → Site1
    - ES2: (PAY ⋈ EMP) → Site4
    - ES3: PROJ → Site 4

$$\begin{aligned} Tot\_cost &= cost(\text{PAYSite} \rightarrow 1) + \\ &\quad cost((\text{PAY} \bowtie \text{EMP}) \rightarrow \text{Site4}) + \\ &\quad cost(\text{PROJ} \rightarrow \text{Site4}) \\ &= 4 + 8 + 1 = 13 \end{aligned}$$

- Both alternatives are not better than ES0, so keep ES0 (or take alternative 2 which has the same cost)

Hill-Climbing Algorithm/6

- **Problems**
  - Greedy algorithm determines an initial feasible solution and iteratively improves it
  - If there are local minima, it may not find the global minimum
  - An optimal schedule with a high initial cost would not be found, since it won't be chosen as the initial feasible solution
- **Example:** A better schedule is
  - PROJ$\rightarrow$Site 4
  - ASG' = (PROJ$\bowtie$ASG)$\rightarrow$Site 1
  - (ASG'$\bowtie$EMP)$\rightarrow$Site 2
  - Total cost$= 1 + 2 + 2 = 5$

SDD-1

- The SDD-1 algorithm extends the hill climbing algorithm with semijoins and has the following properties:
  - Considers semijoins
    - $cost(R \ltimes_A S) = C_{MSG} + size(\pi_A(S)) * C_{TR}$
    - $benefit(R \ltimes_A S) = (1 - SF_{\ltimes}(S.A)) * size(R) * C_{TR}$
  - Does not consider replication and fragmentation
  - Cost of transferring the result to the user site from the final result site is not considered
  - Can minimize either total time or response time
- The SDD-1 algorithm works with and updates a *database profile*:

R	$size(R)$
R1	1500
R2	3000
R3	2000

A	$SF_{\ltimes}$	$size(\pi_A)$
R1.A	0.3	36
R2.A	0.8	320
R2.B	1.0	400
R3.B	0.4	80

SDD-1 Algorithm

Step 1 Include all local processing in the execution strategy ES.

Step 2 Update database profile with effects of local processing.

Step 3 Determine beneficial $\ltimes$, i.e., $cost(\ltimes_i) < benefit(\ltimes_i)$.

Step 4 Remove the most beneficial $\ltimes$ and append it to ES.

Step 5 Update the database profile.

Step 6 Update the set of beneficial semijoins; possibly include new ones.

Step 7 If there are beneficial semijoins go back to Step 4.

Step 8 Find the site where the largest amount of data resides and select it as the result site.

Step 9 For each Ri at the result site, remove semijoins of the form $Ri \ltimes Rj$ where the total cost of ES without this semijoin is smaller than the cost with it.

Step 10 Permute the order of semijoins if doing so would improve the total cost of ES.

Conclusion

- Distributed query optimization is more complex that centralized query processing, since
  - bushy query trees are not necessarily a bad choice
  - one needs to decide what, where, and how to ship the relations between the sites
- Query optimization searches the optimal query plan (tree)
- For N relations, there are $O(N!)$ equivalent join trees. To cope with the complexity heuristics and/or restricted types of trees are considered.
- There are two main strategies in query optimization: randomized and deterministic.
- Semi-joins can be used to implement a join. The semi-joins require more operations to perform, but the data transfer rate is reduced.
- INGRES, System R and Hill Climbing are distributed query optimization algorithms.