More on Query Evaluation

PS#4 will be out tonight.

Relational Database Operations

- We will consider in more detail how to implement:
  - *Selection* (WHERE) Selects rows from table.
  - *Projection* (SELECT) Chose output columns from table.
  - *Join* (explicit JOIN or implicit) Combine two or more tables.
  - *Aggregation* (SUM, MIN, etc.) and GROUP BY
- Since each *op* returns a relation, *ops* can be *composed*! After we cover the operations, we will discuss how to *optimize* queries formed by composing them.

Running Example

- Schema

Sailors (<u>*sid*: integer</u>, *sname*: string, *rating*: integer, *age*: real)
Reserves (<u>*sid*: integer, *bid*: integer, *day*: date</u>, *cardno*: string)

- ~100,000 Reserves:
  - Each tuple is 40 bytes, 100 tuples per page, 1000 pages.
- ~40,000 Sailors:
  - Each tuple is 50 bytes, 80 tuples per page, 500 pages.

Selection

Find rows that satisfy our query's conditions

```
SELECT   *
FROM     Reserves R
WHERE    R.sid = 1000
```

- No Index, Unsorted Data
  - Scan the entire relation, for Reserves 1000 I/Os
- No Index, Sorted by sid
  - Binary search of Reserves $\log_2 1000 \sim 10$ I/Os
- B^+-Tree Index, Clustered on selection attribute
  - Use index to find smallest tuple satisfying selection, scan forward from there, for Reserves 3 I/Os to find starting point + K Blocks having sid=1000 (K ~ 1-2 if resevations ~100 (1%))
- B^+-Tree Index, Unclustered
  - Discussion follows

Using an Index for Selections

- Cost depends on #qualifying tuples, and clustering.
  - Cost of finding qualifying data entries is typically small, but the cost of retrieving records could be large w/o clustering.
  - Example, assuming uniform distribution of ratings (1-10), about 10% of tuples qualify (100 pages, 10000 tuples). With a clustered index, cost is little more than 100 I/Os; if unclustered, could be upto 10000 I/Os!
- *Important refinement for unclustered indexes*:
  1. Find qualifying data entries in index.
  2. Find ***distinct*** *rids* of the pages to be retrieved. (2 ways)
     A. Sort by *rid* while removing replicates
     B. Hash *rids* while eliminating replicates
  3. Scan surviving *rids* while applying selection (result set will be unordered).
  - Ensures each page is considered just once (though # of pages is still likely higher than with clustering).

General Selections

- Selections typically involve more than one attribute with logical conjuncts (and, or)
- Recall we transform to sum-of-product form
- Can be sorted or clustered by only one attribute
- Only a subset of attributes might have indices
- What order to process selection terms?
- How *selective* is a selection term?
  - sid = 1000 — < 1 of 40,000 Sailors
  - age < 20 — ~ 10% of Sailors
  - Rating > 7 — ~ 30 % Sailors

Two Approaches to General Selections

- First approach: Find the *most selective access path*, retrieve tuples using it, and then apply remaining selection terms during scan:
  - *Most selective access path:* An index or file scan that we estimate will require the *fewest page I/Os*.
  - Terms that match this index reduce the number of tuples *retrieved*; other terms are used further discard retrieved tuples, but do not affect number of pages fetched.
  - Consider *day < 8/9/94 AND bid=5 AND sid=3*.
    - A B+ tree index on *day* can be used; then, *bid=5* and *sid=3* must be checked for each retrieved tuple.
    - A hash index on <*bid, sid*> could be used; *day<8/9/94* must then be checked.

Set Operation on Rids

- <u>Second approach</u> (if we have 2 or more matching indexes):
  - Use indexes to get sets of data records pids using each matching index.
  - *Intersect* and/or *union* these sets of rids
    Retrieve the pages with records and apply tests.
  - Consider (*day<8/9/94 AND bid=5 AND sid=3)*.
    - If we have a B^+ tree index on *day* and an index on *sid*, both unclustered, we can retrieve *distinct pids* satisfying *day<8/9/94 AND sid=3.*
    - If we have a Hash index on *(sid,bid)* we can use it to extract the *pids* of records satisfying *bid=5 AND sid=3*
    - Intersect these *pid sets*, then retrieve all records and check.

Projection

```
SELECT    DISTINCT R.sid, R.bid
FROM      Reserves R
```

- Modified external sorting:
  - Modify Pass 0 of external sort to eliminate unneeded fields. Thus, writing out fewer pages. Tuples merged in subsequent sprting passes are smaller than tuples of the original relation. (i.e. Instead of 40 bytes/record, perhaps 8, so 500 can fit in a page. Size ratio depends on # and size of fields that are dropped.)
  - Modify merge passes to eliminate duplicates. Thus, number of result tuples is even smaller than input. (Depends on # of duplicates.)
  - Cost: In Pass 0, reads all original pages, but writes out fewer pages (same number of smaller tuples). In merge passes, fewer tuples are written out due to the eliminated duplicates.

Projection Based on Hashing

- Modified hashing:
  - *Partitioning phase*: Read R using one input buffer. For each tuple, discard unwanted fields, apply hash function h_1 to direct output to one of B-1 output buffers (hash buckets).
    - Result is B-1 partitions (of tuples with no unwanted fields). Tuples in different partitions are guaranteed to be distinct.
  - *Duplicate elimination phase*: Foreach partition either:
    - Build another "in-memory" hash table, using hash function h_2 ($\neq h_1$), that discards duplicates (handled on collisions).
    - Maintian sorted partitions while inserting to eliminate duplicates
  - Cost: For partitioning, read R, write out each tuple, but with fewer fields. This is read in next phase.

Discussion of Projection

- Sort-based approach is the standard; better handles skewed attribute distributions and result is sorted.
- If an index on the relation contains the wanted projection attributes as its search key, then we can use an *index-only* scan (no fetching of the primary data pages).
- If an ordered (i.e., tree) index contains all wanted attributes as a *prefix* of its search key's we can
  - Retrieve data entries in order (index-only scan), discard unwanted fields, compare adjacent tuples to check for duplicates.

Equijoins w/one common column

```
SELECT    *
FROM      Reserves R, Sailors S
WHERE     R.sid=S.sid
```

- Implicit JOINS are very common! Must be carefully optimized. Often R×S is very large; so, R×S followed by a selection is inefficient.
- Assume: M tuples in R, p_R tuples/page,
  N tuples in S, p_S tuples/page.
- *Cost metric*: # of I/Os. We will ignore output costs.

Basic Nested Loops Join

```
foreach tuple r in R:
   foreach tuple s in S:
      if r_i == s_j :
         add <r, s> to result
```

- Naïve Approach: For each tuple in the *outer* relation R, we scan the entire *inner* relation S (i.e. R × S).
  - Cost: $M + (p_R * M) * N = 1000 + 100*1000*500$ I/Os.
- ***Page-at-a-time*** Nested Loops join: For each *page* of R, get each *page* of S, and handle all matching pairs of tuples <r, s>, where r is in R-page and S is in S-page.
  - Cost: $M + M*N = 1000 + 1000*500$ I/Os
  - If smaller relation (S) is outer, cost = 500 + 500*1000 I/Os

Index Nested Loops (INL) Join

```
foreach tuple r in R:
    foreach tuple s in S where r_i == s_j:
        add <r, s> to result
```

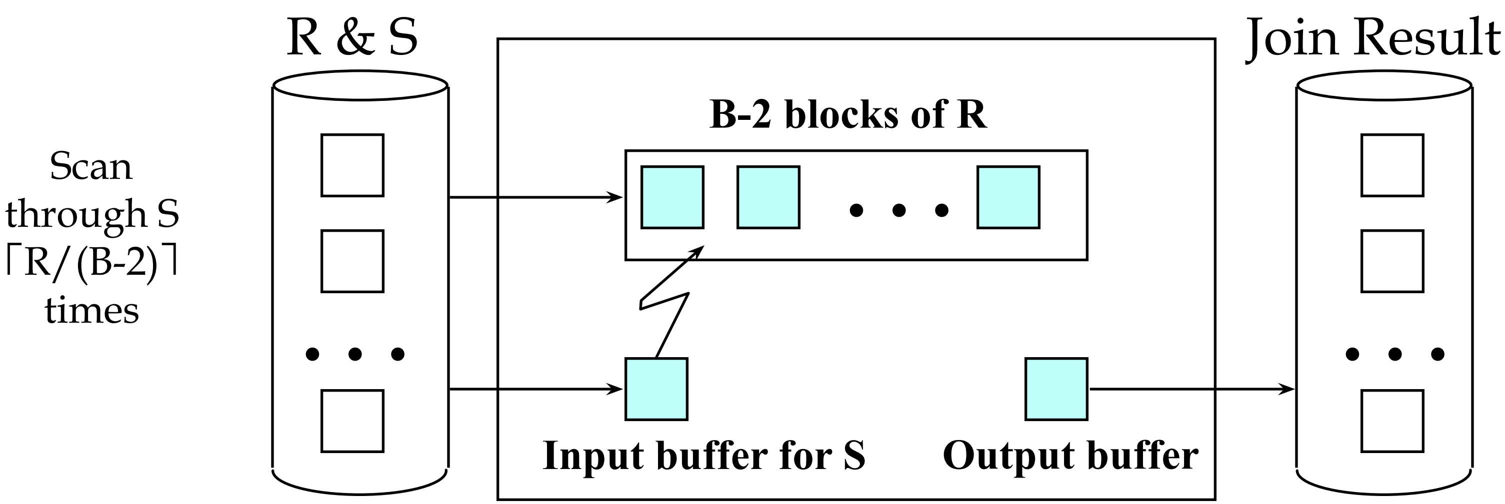
- If there is an index on the join column of one relation (say S), make it the inner loop, and exploit the index.
  - Cost: M + ((M*p_R) * cost of finding matching S tuples)
- For each R tuple, cost of probing S index is about 1.2 page reads for a hash index, and 2-4 for B+ tree. Cost of then finding actual S tuples depends on clustering.
  - Clustered index: 1 I/O (typical), unclustered: upto 1 I/O per matching S tuple.

Examples of Index Nested Loops

- Hash-index (Alt. 2) on *sid* of Sailors (as inner):
  - Scan Reserves: 1000 page I/Os, to retrieve 100*1000 tuples.
  - For each Reserves tuple: 1.2 I/Os to get *pid* from index, plus 1 I/O to get (exactly one) matching Sailors tuple. Total: 1000 + 2.2 * 100,000 = 221,000 I/Os.
- Hash-index (Alt. 2) on *sid* of Reserves (as inner):
  - Scan Sailors: 500 page I/Os, to retrieve 80*500 tuples.
  - For each Sailors tuple: 1.2 I/Os to find index page with *sid* search key, plus cost of retrieving, possibly multiple, matching Reserves tuples. Assuming uniform distribution, 2.5 reservations per sailor (100,000 / 40,000). Cost of retrieving them is 1 or 2.5 I/Os depending on whether the index is clustered. Total = 500 + (1.2 + 2) * 40,000 = 128,500 I/Os

Block Nested Loops (BNL) Join

- Small twist on Simple Nested Loops
- Use one page as an input buffer for scanning the inner loops relation, S, one page as the output buffer, and use all remaining (B-2) pages to hold a "block" of pages from the outer loops relation, R.
  - For each matching tuple r in R-block, s in S-page, add <r, s> to result. Then read next R-block, scan S, etc.



Examples of Block Nested Loops

- Cost: $M + N \lceil M/(B-2)$
- With Reserves (R) as outer and 102 buffer pages:
  - Cost of scanning R is *M = 1000* I/Os over 10 *passes*.
  - Per pass of R, we scan Sailors (S); 10*500 I/Os.
  - With space for 90 pages of R, we scan S 12 times.
- With 100-page block of Sailors as outer:
  - Cost of scanning S is *M = 500* I/Os over 5 *passes*.
  - Per pass of S, we scan Reserves (R); 5*1000 I/Os.
- Better yet, we can double buffer the inner loop with a pass size of (B-3), allowing us to simultaneously fetch the next block while joining current one

Sort-Merge Join (SMJ) Review

- Sort R and S on the join column, then scan them while merging on the join col.) and outputing result tuples.
  - Advance scan of R until current R-tuple >= current S tuple, then advance scan of S until current S-tuple >= current R tuple; do this until current R tuple = current S tuple.
  - At this point, one-or-more, ρ, R tuples match one-or-more, σ, S tuples; output <r, s> for all pairs of such tuples ($\rho\times\sigma$).
  - Then resume scanning R and S.
- Cost: M log M + N log N + (M+N)

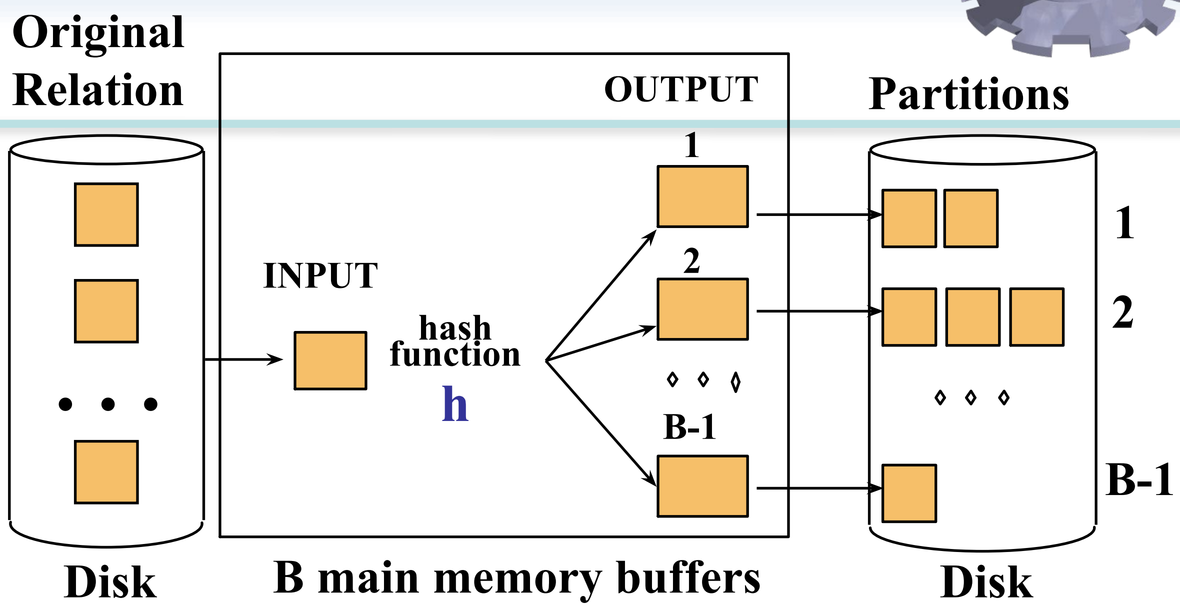
Refinements of Sort-Merge Join

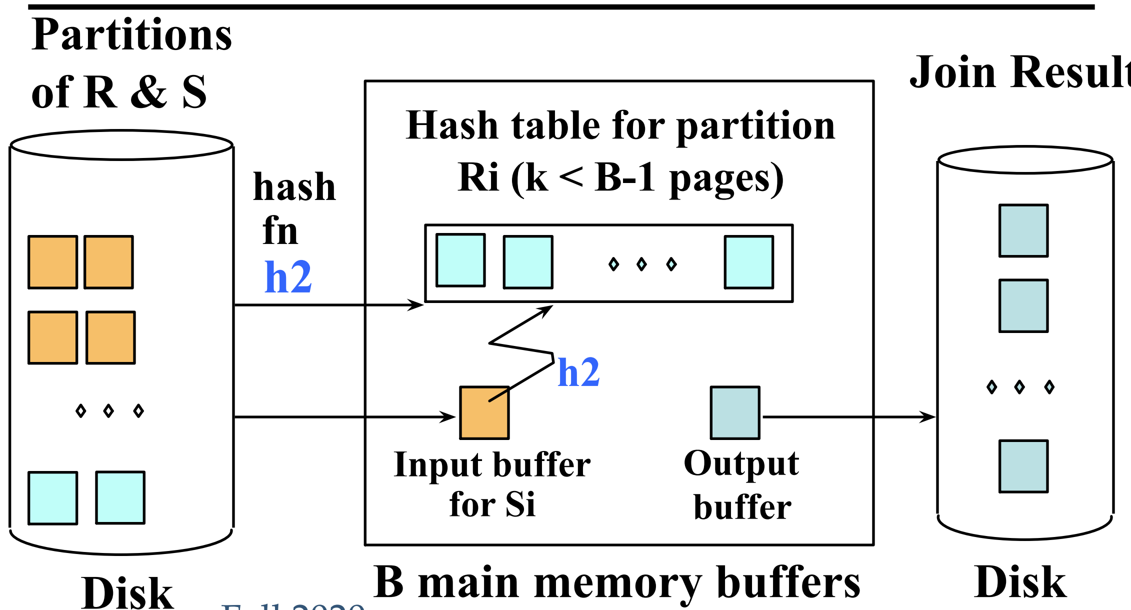
- Combine the merging phases of *external sorting* of R and S with the merging required for the join.
  - Using the sorting refinement that merges multiple runs each pass, we sort R and S up to their last merge pass.
  - Allocate 1 page per run of each relation, and "merge" while checking the join condition.
  - Cost: read+writes in (Pass 0.. Pass N-1) + read each relation in (only) merging pass (+ writing of result tuples).
  - Typically reduces I/O cost by a factor of ½.
- In practice, cost of sort-merge join, like the cost of external sorting, is nearly *linear*.

Hash-Join

- Partition both relations using a common hash function, h, (R tuples in partition i will only match S tuples in partition i).

- Read in a partition of R, hash it using h2 (<> h). Scan matching partition of S, search for matches.





Observations on Hash-Join

- We want each partition of R to fit in B-2 buffer pages, so #partitions, k = M / (B – 2), if we assume no skew
- If we build an in-memory hash table to speed up the matching of tuples, a little more memory is needed.
- If the hash function does not partition uniformly, one or more R partitions may not fit in memory. Can apply hash-join technique recursively to this partition and do the join of this R-partition with corresponding S-partition.

Cost of Hash-Join

- In partitioning phase, read+write both relns; 2(M+N). In matching phase, read both relns; M+N I/Os.
- In our running example, this is a total of 4500 I/Os.
- Sort-Merge Join vs. Hash Join:
  - Both have a cost of 3(M+N) I/Os. Hash-Join is superior if relation sizes differ greatly. Also, Hash-Join shown to be highly parallelizable.
  - Sort-Merge insensitive to data skew; and result is sorted.

Aggregate Operations (AVG, MIN, etc.)

- **Without grouping:**
  - In general, requires scanning the relation.
  - Given index whose search key includes all attributes in the SELECT or WHERE clauses, can do index-only scan.
- **With grouping:**
  - Sort on group-by attributes, then scan relation and compute aggregate for each group. (Can improve upon this by combining sorting and aggregate computation.)
  - Similar approach based on hashing on group-by attributes.
  - Given tree index whose search key includes all attributes in SELECT, WHERE and GROUP BY clauses, can do index-only scan; if group-by attributes form prefix of search key, can retrieve data entries/tuples in group-by order.

Summary

- A virtue of relational DBMSs: *queries are composed of a few basic operators*; the implementation of these operators can be carefully tuned (and it is important to do this!).
- Alternative implementations for each operator; no universally superior technique for most operators.
- Must consider available alternatives for each operation in a query and choose best one based on system statistics, etc. This is part of the broader task of optimizing a query composed of several ops.