Database Management Systems

Lecture 6

Evaluating Relational Operators

Query Optimization

* queries – composed of <u>relational operators</u>*:

- selection (σ)
  - selects a subset of records from a relation
- projection (π)
  - eliminates certain columns from a relation
- join ($\otimes$)
  - combines data from two relations
- cross-product (R1 $\times$ R2)
  - returns every record in R1 concatenated with every record in R2
- set-difference (R1 $-$ R2)
  - returns records that belong to R1 and don't belong to R2
- union (R1 $\cup$ R2)
  - returns all records in relations R1 and R2

*Review *Relational Algebra* - lecture notes (*Databases* course)

* queries – composed of <u>relational operators</u>:

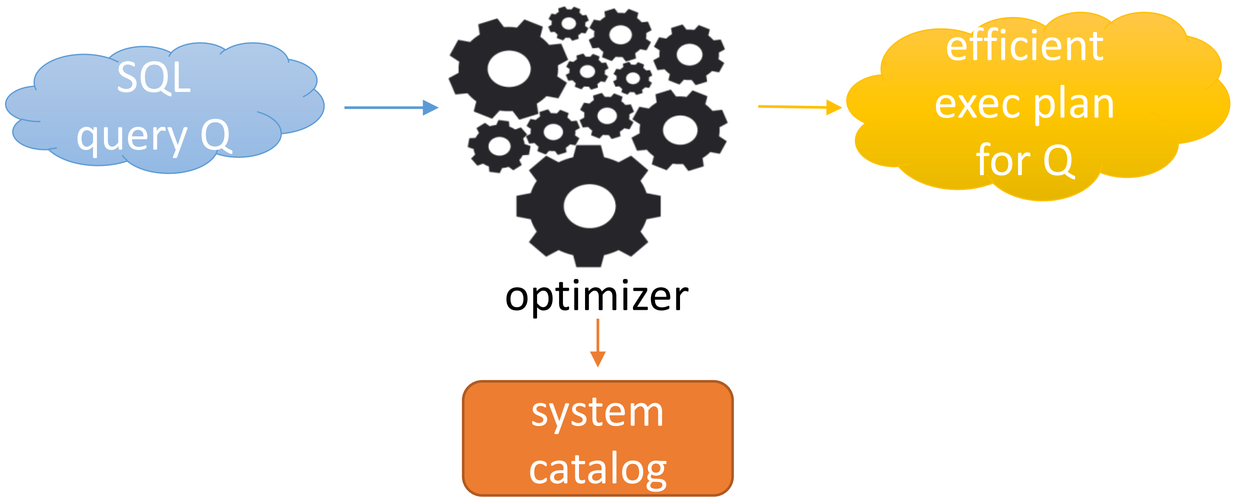
- intersection (R1 ∩ R2)
  - returns records that belong to both R1 and R2
- grouping and aggregate operators (algebra extensions)

- every operation returns a relation => operations can be composed

- each operator has several implementation algorithms

* optimizer

- input: SQL query Q
- output: an efficient execution plan for evaluating Q



- takes into account the manner in which data is stored (such information is available in the system catalog)
- chooses an implementation for each operator in the query, taking into account the size of relations, the size of available buffer pool; the existence of indexes and sort orders; the buffer replacement policy
- determines an application order for the operators in the query

* the algorithms for various operators are based on 3 techniques:

- <u>iteration</u>:
  - examine iteratively:
    - all tuples in input relations

  or
    - data entries in indexes, provided they contain all the necessary fields (data entries are smaller than data records)
- <u>indexing</u>:
  - used when the query contains a selection condition or a join condition
  - examine only the tuples that meet the condition, using an index
- <u>partitioning</u>:
  - partition the tuples
  - decompose operation into collection of cheaper operations on partitions

* the algorithms for various operators are based on 3 techniques:

- partitioning:
    - partitioning techniques
        - sorting
        - hashing

* <u>access paths</u>

- *access path* = way of retrieving tuples from a relation; 2 possibilities:
  - file scan

  or
  - an index *I* + a matching selection condition *C*
- condition *C* matches index *I* if *I* can be used to retrieve just the tuples* satisfying *C*
- if relation *R* has an index *I* that matches selection condition *C*, then there are at least 2 access paths for *R* (scan; index + condition)

*Review *Indexes* - lecture notes (*Databases* course)

* <u>access paths</u> - example:

- relation *Students[SID, Name, City]*
- there's a tree index *I* on *Students* with search key <*Name*>
- query Q:

```
SELECT *
FROM Students
WHERE Name = 'Ionescu'
```

- condition *C*: *Name = 'Ionescu'*
- *C* matches *I*, i.e., index *I* can be used to retrieve only the *Students* tuples satisfying *C*
- the same would be true for, say, *C*: *Name > 'Ionescu'*

* <u>access paths</u> - example:

- relation *Students[SID, Name, City]*
- there's a hash index *I* on *Students* with search key <*Name*>
- query Q:

  SELECT *

  FROM Students

  WHERE Name = 'Ionescu'
- condition *C*: *Name = 'Ionescu'*
- *C* matches *I*, i.e., index *I* can be used to retrieve only the *Students* tuples satisfying *C*
- however, condition *Name > 'Ionescu'* doesn't match *I* (since *I* is a hash index; it cannot be used to retrieve just the tuples where *Name* > '*Ionescu*')

* <u>access paths</u>

- to sum up:
  - condition *C: attr op value*, $op \in \{<, <=, =, <>, >=, >\}$
  - condition *C* matches index *I* if:
  - the search key of *I* is *attr* and:
    - *I* is a tree index or
    - *I* is a hash index and *op* is *=*

* <u>access paths</u>

- <u>selectivity</u> of an access path
  - the number of retrieved pages when using the access path to obtain the desired tuples
  - both data pages and index pages are counted
- example:

```
SELECT *
FROM Students
WHERE Name = 'Ionescu'
```

  - there are 2 access paths for *Students*:
    - file scan – selectivity could be 1000
    - matching index I with search key <Name> – selectivity could be 3
- <u>most selective</u> access path
  - retrieves the fewest pages, i.e., data retrieval cost is minimized

* <u>general selection conditions</u>

- previous examples included only simple conditions of the form *attr op value*
- in general, a selection condition can contain one or several terms of the form:
  - *attr op constant*
  - *attr1 op attr2*,

  combined with ∧ and ∨

```
SELECT *
FROM Exams
WHERE SID = 7 AND EDate = '04-01-2019'
```

$$\sigma_{SID=7 \wedge EDate='04-01-2019'}(Exams)$$

* general selection conditions

- CNF – conjunctive normal form
  - standard form for general selection conditions
  - condition in CNF:
    - collection of conjuncts connected with the ∧ operator
    - a *conjunct* has one or more terms connected with the ∨ operator
    - *term*:
      - *attr op constant*
      - *attr1 op attr2*

- example:

  condition *(EDate < '4-1-2019' ∧ Grade = 10) ∨ CID = 5 ∨ SID = 3*

  is rewritten in CNF:

  (EDate < '4-1-2019' ∨ CID = 5 ∨ SID = 3) ∧ (Grade = 10 ∨ CID = 5 ∨ SID = 3)

* <u>general selection conditions</u> matching an index

- index I with search key <a, b, c>, relation R[a, b, c, d, e]

Condition	B+ tree index	Hash index
a = 10 AND b = 5 AND c = 2	Yes	Yes
a = 10 AND b = 5	Yes	No
b = 5	No	No
b = 5 AND c = 2	No	No
d = 2	No	No
a = 20 AND b = 10 AND c = 5 AND d = 11	Yes (partly)	Yes (partly)

Condition – CNF selection condition
B+ tree index / Hash index – *B+ tree / hash index* I matches (*Yes*) / doesn't match (*No*) / matches a part of (*Yes (partly)*) the selection condition

- for the condition in the last row (a = 20 AND b = 10 AND c = 5 AND d = 11):
  - use index I to retrieve tuples satisfying *a = 20 AND b = 10 AND c = 5*, then apply *d = 11* to each retrieved tuple

* <u>general selection conditions</u> matching an index

- index I1 with search key <a, b>
- B+ tree index I2 with search key <c>
- relation R[a, b, c, d]

Condition	Indexes
c < 100 AND a = 3 AND b = 5	- use I1 or I2 to retrieve tuples - then check terms in the selection condition that do not match the index for each retrieved tuple - e.g., use the B+ tree index to retrieve tuples where c < 100; then apply *a = 3 AND b = 5* to each retrieved tuple

* <u>general selection conditions</u> matching an index

- index *I*, general selection condition *C* (CNF)

- hash index *I*
- condition *C* of the form:
  - $\bigwedge_{i=1}^{n} T_i$
  - term T_i: $attr = value$

  (there are no disjunctions in *C*)
- *I* matches *C* if *C* contains exactly one term for each attribute in the search key of *I*

Condition	Hash index with search key <a, b, c>
a = 10 AND b = 5 AND c = 2	Yes
a = 10 AND b = 5	No
b = 5	No
b = 5 AND c = 2	No

* <u>general selection conditions</u> matching an index

- index *I*, general selection condition *C* (CNF)

- tree index *I*
- condition *C* of the form:
  - $\bigwedge_{i=1}^{n} T_i$
  - term T_i: $attr\ op\ value$; *op* ϵ {<, <=, =, <>, >=, >}
- *I* matches *C* if *C* contains exactly one term for each attribute in a prefix of the search key of *I*
- examples of prefixes for search key <a, b, c, d>: <a>, <a, b, c>; <a, c> and <b, c>, on the other hand, are not prefixes for this search key

Condition	B+ tree index with search key <a, b, c>
a = 10 AND b = 5 AND c = 2	Yes
a = 10 AND b = 5	Yes
b = 5	No
b = 5 AND c = 2	No

* running example - schema

- Students (<u>SID: integer</u>, SName: string, Age: integer)
- Courses (<u>CID: integer</u>, CName: string, Description: string)
- Exams (<u>SID: integer, CID: integer</u>, EDate: date, Grade: integer, FacultyMember: string)

- Students
  - every record has 50 bytes
  - there are 80 records / page
  - 500 pages of Students tuples
- Courses
  - every record has 50 bytes
  - there are 80 records / page
  - 100 pages of Courses tuples

* running example - schema

- Students (SID: integer, SName: string, Age: integer)
- Courses (CID: integer, CName: string, Description: string)
- Exams (SID: integer, CID: integer, EDate: date, Grade: integer, FacultyMember: string)

- Exams
  - every record has 40 bytes
  - there are 100 records / page
  - 1000 pages of Exams tuples

* <u>joins</u>

```
SELECT *
FROM Exams E, Students S
WHERE E.SID = S.SID
```

- algebra: $E \otimes_{E.SID=S.SID} S$
  - to be carefully optimized
  - size of $E \times S$ is large, so computing $E \times S$ followed by selection is inefficient
- E
  - M pages
  - p_E records / page
- S
  - N pages
  - p_S records / page
- evaluation: number of I/O operations

* joins – implementation techniques
  - iteration
    - Simple/Page-Oriented Nested Loops Join
    - Block Nested Loops Join
  - indexing
    - Index Nested Loops Join
  - partitioning
    - Sort-Merge Join
    - Hash Join

- equality join, one join column
  - join condition: $E_i = S_j$

Simple Nested Loops Join

```
foreach tuple e ∈ E do
    foreach tuple s ∈ S do
        if e_i == s_j then add <e, s> to the result
```

- for each record in the outer relation E, scan the entire inner relation S
- cost
  - $\mathbf{M + p_E * M * N}$ = 1000 + 100*1000*500 I/Os = 1000 + (5 * 10^7) I/Os
    - M I/Os – cost of scanning E
    - N I/Os – cost of scanning S
    - S is scanned p_E* M times (there are p_E* M records in the outer relation E)

* E - M pages, p_E records / page *	* 1000 pages *	* 100 records / page*
* S - N pages, p_S records / page *	* 500 pages *	* 80 records / page *

Page-Oriented Nested Loops Join

```
foreach page pe ∈ E do
    foreach page ps ∈ S do
        if e_i == s_j then add <e, s> to the result
```

- for each page in E read each page in S
- pairs of records <e, s> that meet the join condition are added to the result (where record *e* is on page *pe*, and record *s* – on page *ps*)
- refinement of Simple Nested Loops Join

Page-Oriented Nested Loops Join

```
foreach page pe ∈ E do
    foreach page ps ∈ S do
        if e_i == s_j then add <e, s> to the result
```

- cost
  - **M + M*N** = 1000 + 1000*500 I/Os = 501.000 I/Os
    - M I/Os – cost of scanning E; N I/Os – cost of scanning S
    - S is scanned M times
    - significantly lower than the cost of Simple Nested Loops Join (improvement - factor of p_E)
  - if the smaller table (*S*) is chosen as outer table:

  => cost = 500 + 500 * 1000 I/Os = 500.500 I/Os

* E - M pages, p_E records / page *   * 1000 pages *   * 100 records / page*

* S - N pages, p_S records / page *   * 500 pages *   * 80 records / page *

Block Nested Loops Join

- previously presented join algorithms do not use buffer pages effectively
- join relations R1 and R2; R1 – the smaller relation
- assumption – the smaller relation fits in main memory
- improvement:
  - store smaller relation R1 in memory
  - keep at least 2 extra buffer pages B1 and B2
  - use B1 to read the larger relation R2 (one page at a time)
  - use B2 as the output buffer (i.e., for tuples in the result of the join)
  - for each tuple in R2, search R1 for matching tuples

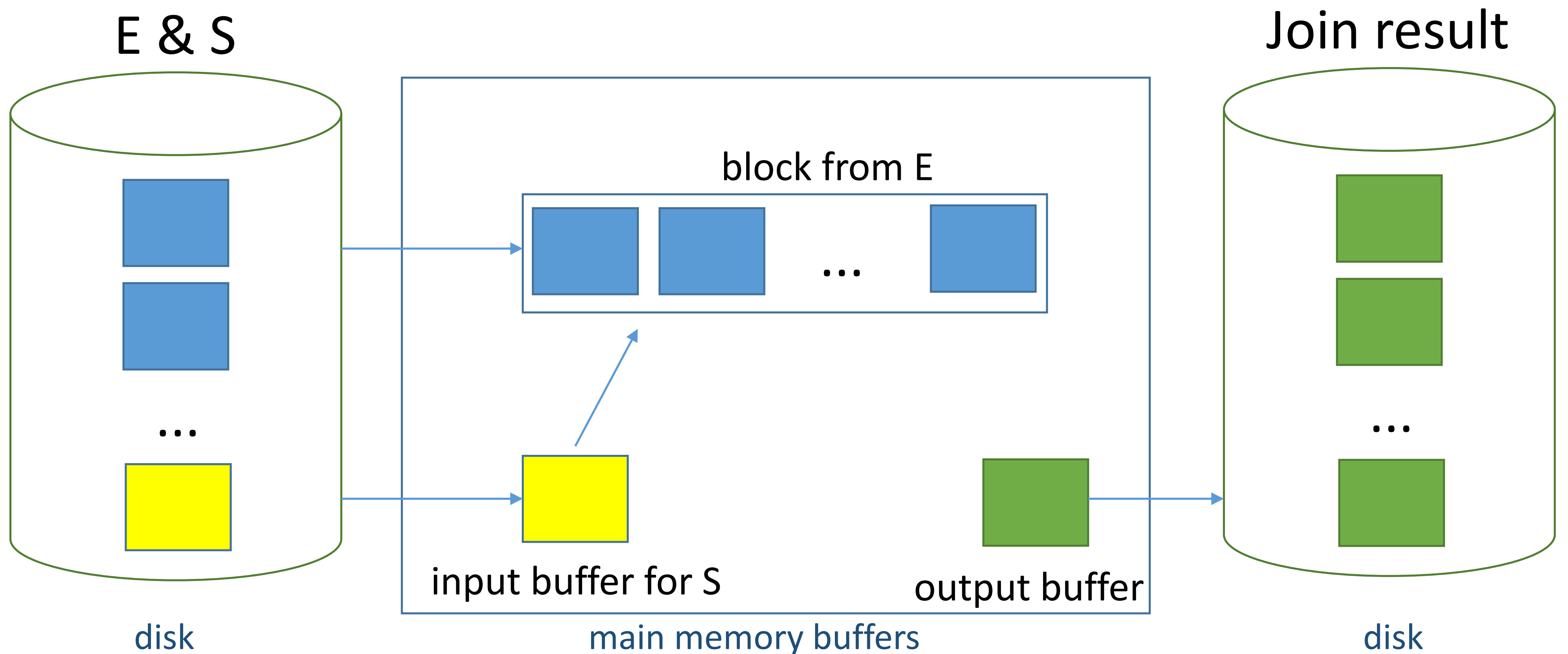
  => optimal cost: *number of pages in R1 + number of pages in R2*, since R1 is scanned only once, R2 is also scanned only once

Block Nested Loops Join

- refinement
  - don’t store the smaller relation in main memory as is, build an in-memory hash table for it instead
  - the I/O cost remains unchanged, but the CPU cost is usually much lower (since for each tuple in the larger relation, the smaller relation is examined to find matching tuples)

Block Nested Loops Join

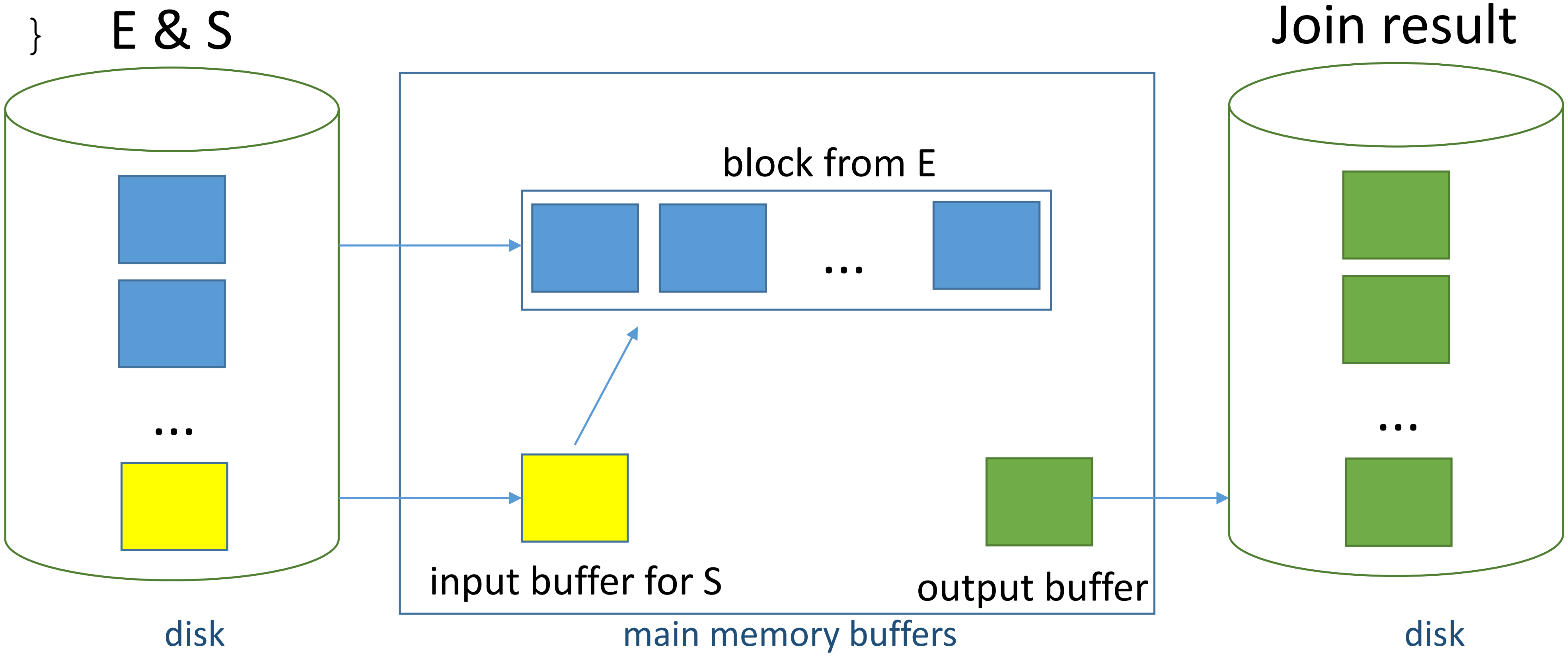
- if there isn't enough main memory to hold one of the input relations:
  - use one buffer page to scan the inner table (e.g., S)
  - use one page for the result
  - use all remaining pages to read a *block* from the outer table (e.g., E)
    - block – set of pages from E that fit in main memory



Block Nested Loops Join

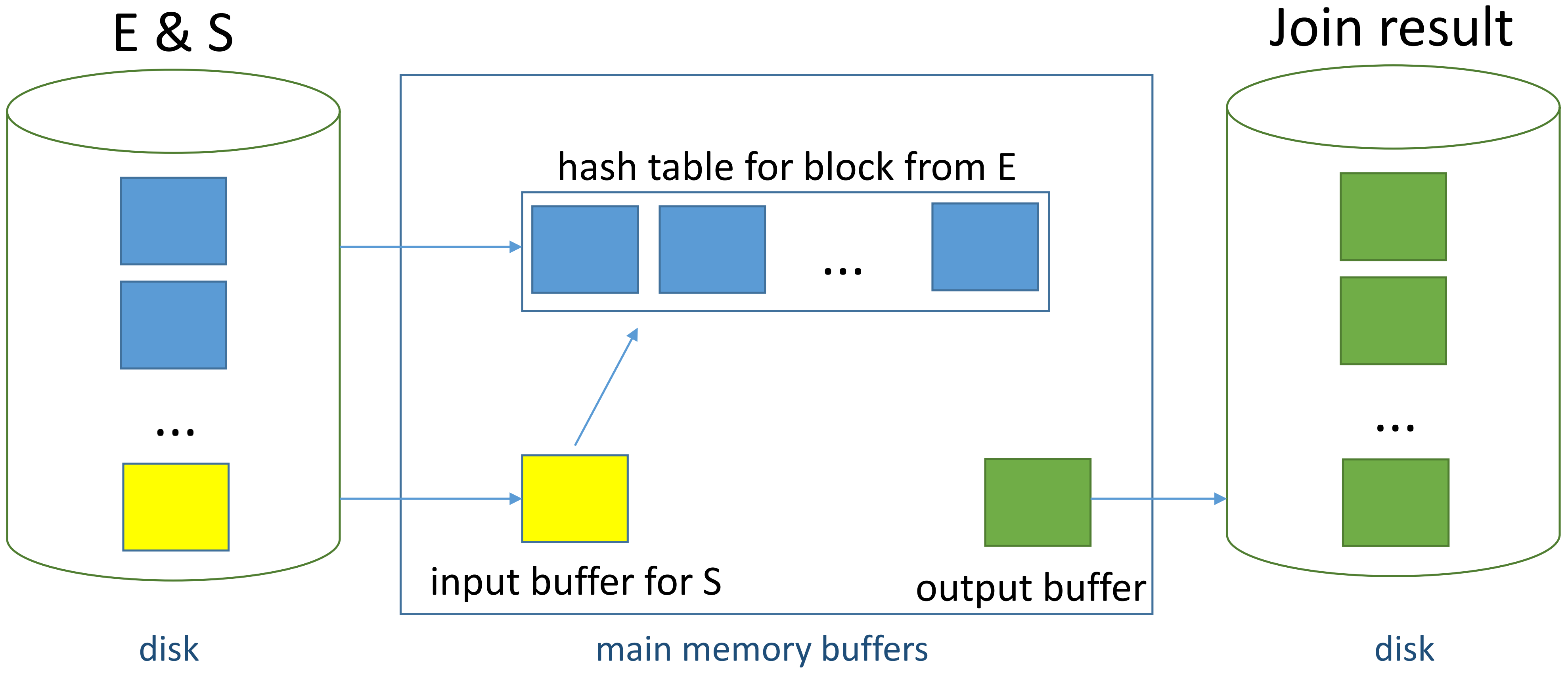
```
foreach block be ∈ E do
    foreach page ps ∈ S do
      {
          for all pairs of tuples <e, s> that meet the join
            condition, where e ∈ be and s ∈ ps,
                 add <e, s> to the result
      }
```

- inner relation S is scanned once for each block in outer relation E
- outer relation E is scanned once



Block Nested Loops Join

- refinement to efficiently find matching tuples
  - build main-memory hash table for the block of E
  - trade-off: reduce size of E block



Block Nested Loops Join

- cost
  - scan of outer table + number of blocks in outer table * scan of inner table
  - number of outer blocks = $\left\lceil \frac{\text{number of pages in outer table}}{\text{size of block}} \right\rceil$
  - outer table: Exams (E), a block can hold 100 pages
    - scan cost for E: 1000 I/Os
    - number of blocks: $\left\lceil \frac{1000}{100} \right\rceil$ = 10
    - foreach block in E, scan Students (S): 10*500 I/Os

  => total cost = 1000 + 10 * 500 = **6000 I/Os**

* E - M pages, p_E records / page * * 1000 pages * * 100 records / page*

* S - N pages, p_S records / page * * 500 pages * * 80 records / page *

Block Nested Loops Join

- cost
  - scan of outer table + number of blocks in outer table * scan of inner table
  - number of outer blocks = $\left\lceil \frac{\text{number of pages in outer table}}{\text{size of block}} \right\rceil$
  - outer table: Exams (E)
    - suppose the buffer has 90 pages available for E, i.e., block of 90 pages

    => number of blocks: $\left\lceil \frac{1000}{90} \right\rceil$ = 12

    => S is scanned 12 times
    - scan cost for E: 1000 I/Os
    - foreach block in E, scan Students (S): 12*500 I/Os

  => total cost = 1000 + 12 * 500 = **7000 I/Os**

* E - M pages, p_E records / page *	* 1000 pages *	* 100 records / page*
* S - N pages, p_S records / page *	* 500 pages *	* 80 records / page *

<u>Block Nested Loops Join</u>

- <u>cost</u>
  - scan of outer table + number of blocks in outer table * scan of inner table
  - number of outer blocks = $\left\lceil \frac{\text{number of pages in outer table}}{\text{size of block}} \right\rceil$
  - outer table: Students (S), block of 100 pages
    - scan cost for S: 500 I/Os
    - number of blocks: $\left\lceil \frac{500}{100} \right\rceil$ = 5
    - for each block in S, scan E: 5 * 1000 I/Os

  => total cost = 500 + 5 * 1000 = **5500 I/Os**

* E - M pages, p_E records / page *	* 1000 pages *	* 100 records / page*
* S - N pages, p_S records / page *	* 500 pages *	* 80 records / page *

Index Nested Loops Join

```
foreach tuple e in E do
     foreach tuple s in S where e_i == s_j
          add <e, s> to the result
```

- if there is an index on the join column of S, S can be considered as inner table and the index can be used
- cost
  - M + ((M*p_E) * cost of finding corresponding records in S)

* E - M pages, p_E records / page *   * 1000 pages *   * 100 records / page*

* S - N pages, p_S records / page *   * 500 pages *   * 80 records / page *

Index Nested Loops Join

- for a record e in E:
  - the cost of examining the index on S is:
    - approx. 1.2 for a hash index (typical cost for hash indexes)
    - typically 2-4 for a B+-tree index

  - the cost of reading corresponding records in S:
    - for a clustered index:
      - plus one I/O for each outer tuple in E (typically)
    - for a nonclustered index:
      - up to one I/O for each corresponding record in S
        (worst case – there are *n* matching records in S located on *n* different pages!)

Index Nested Loops Join

- hash index on SID in Students (Students – inner table)
- scan Exams:
  - cost = 1000 I/Os, with a total of 100*1000 records
- for each record in Exams:
  - (on average) 1.2 I/Os to obtain the page in the hash index (i.e., the page containing the rid of the matching Students tuple)
    and
  - 1 I/O to retrieve the page in Students that contains the matching tuple (exactly one! – since SID is a key in Students, i.e., there is one matching Students tuple for an exam)

=> cost to retrieve matching Students tuples: 1000 * 100 * (1.2 + 1) = 220.000

- total cost: 1000 + 220.000 = 221.000 I/Os

* E - M pages, p_E records / page * * 1000 pages * * 100 records / page*

* S - N pages, p_S records / page * * 500 pages * * 80 records / page *

Index Nested Loops Join

- hash index on SID in Exams
- scan Students - 500 I/Os, 80*500 records
- for each record in Students:
    - (on average) 1.2 I/Os to obtain the index page + the cost of reading the matching records in Exams
        - assume exams are uniformly distributed => 2.5 exams for each student (100.000 Exams tuples / 40.000 Students tuples)
        - the cost of retrieving the corresponding records:
            - if the index is clustered: 1 I/O
            - if the index is unclustered: 2.5 I/Os

=> total cost: 500 + 80*500*(1.2 + 1) = 500 + 88.000 = 88.500 I/Os or
total cost: 500 + 80*500*(1.2 + 2.5) = 500 + 148.000 = 148.500 I/Os

* E - M pages, p_E records / page * * 1000 pages * * 100 records / page*

* S - N pages, p_S records / page * * 500 pages * * 80 records / page *

References

- [Ra00] RAMAKRISHNAN, R., GEHRKE, J., Database Management Systems (2nd Edition), McGraw-Hill, 2000
- [Da03] DATE, C.J., An Introduction to Database Systems (8th Edition), Addison-Wesley, 2003
- [Ga08] GARCIA-MOLINA, H., ULLMAN, J., WIDOM, J., Database Systems: The Complete Book, Prentice Hall Press, 2008
- [Ra07] RAMAKRISHNAN, R., GEHRKE, J., Database Management Systems, McGraw-Hill, 2007, http://pages.cs.wisc.edu/~dbbook/openAccess/thirdEdition/slides/slides3ed.html
- [Si10] SILBERSCHATZ, A., KORTH, H., SUDARSHAN, S., Database System Concepts, McGraw-Hill, 2010, http://codex.cs.yale.edu/avi/db-book/
- [Ul11] ULLMAN, J., WIDOM, J., A First Course in Database Systems, http://infolab.stanford.edu/~ullman/fcdb.html