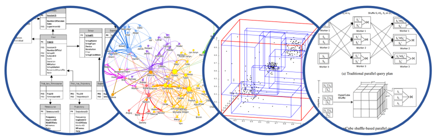


Database System Internals

Operator Algorithms (part 2)

Paul G. Allen School of Computer Science and Engineering
University of Washington, Seattle

1

Announcements

- Homework 2 released
  - Due January 31st
- 544 paper 1 report due Today
- Lab 2 posted after class
  - Part 1 (operator algos) due Friday
  - Part 2 (insert/delete support) due following Friday

2

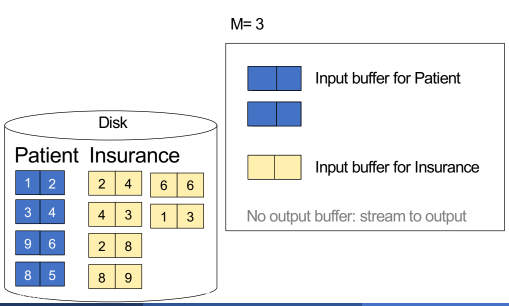
Block-Memory Refinement

```
for each group of M-1 pages r in R do
  for each page of tuples s in S do
    for all pairs of tuples t1 in r, t2 in s
      if t1 and t2 join then output (t1,t2)
```

What is the Cost?

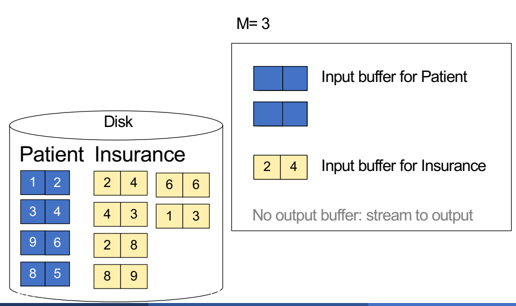
3

Block Memory Refinement



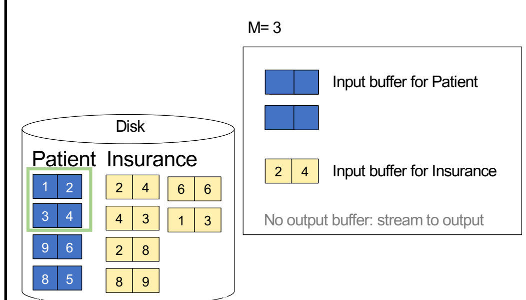
4

Block Memory Refinement



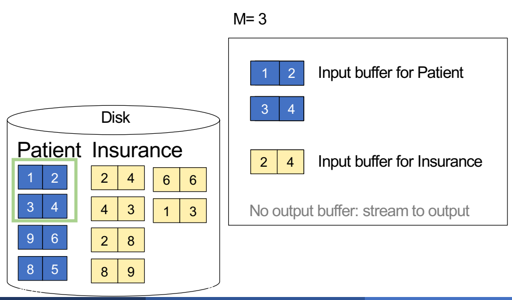
5

Block Memory Refinement



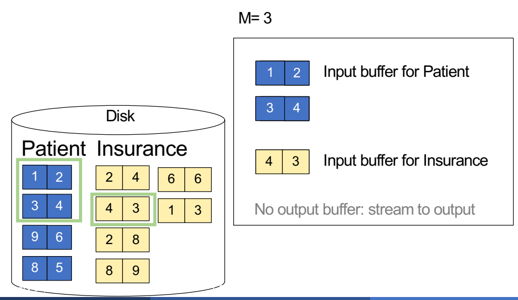
6

Block Memory Refinement



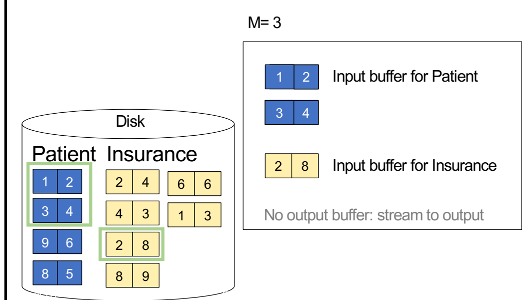
7

Block Memory Refinement



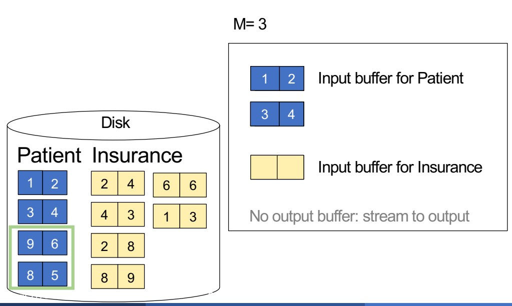
8

Block Memory Refinement



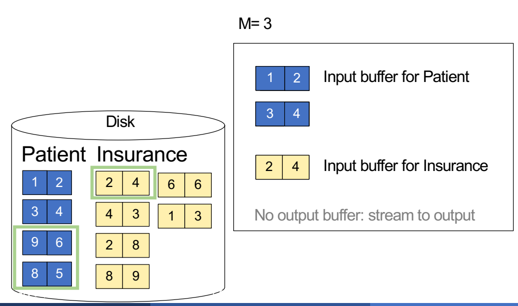
9

Block Memory Refinement



10

Block Memory Refinement



11

Block Memory Refinement

```
for each group of M-1 pages r in R do
  for each page of tuples s in S do
    for all pairs of tuples t1 in r, t2 in s
      if t1 and t2 join then output (t1,t2)
```

What is the Cost?

12

Block Memory Refinement

```
for each group of M-1 pages r in R do
  for each page of tuples s in S do
    for all pairs of tuples t1 in r, t2 in s
      if t1 and t2 join then output (t1,t2)
```

- Cost: B(R) + B(R)B(S)/(M-1)

What is the Cost?

13

Outline

- **Join operator algorithms**
  - One-pass algorithms (Sec. 15.2 and 15.3)
  - Index-based algorithms (Sec 15.6)
  - Two-pass algorithms (Sec 15.4 and 15.5)

19

Index Based Selection

Selection on equality: $\sigma_{a=v}(R)$

- B(R)= size of R in blocks
- T(R) = number of tuples in R
- V(R, a) = # of distinct values of attribute a

20

Index Based Selection

Selection on equality: $\sigma_{a=v}(R)$

- B(R)= size of R in blocks
- T(R) = number of tuples in R
- V(R, a) = # of distinct values of attribute a

What is the cost in each case?

- Clustered index on a:
- Unclustered index on a:

21

Index Based Selection

Selection on equality: $\sigma_{a=v}(R)$

- B(R)= size of R in blocks
- T(R) = number of tuples in R
- V(R, a) = # of distinct values of attribute a

What is the cost in each case?

- Clustered index on a: B(R)/V(R,a)
- Unclustered index on a:

22

Index Based Selection

Selection on equality: $\sigma_{a=v}(R)$

- B(R)= size of R in blocks
- T(R) = number of tuples in R
- V(R, a) = # of distinct values of attribute a

What is the cost in each case?

- Clustered index on a: B(R)/V(R,a)
- Unclustered index on a: T(R)/V(R,a)

23

Index Based Selection

Selection on equality: $\sigma_{a=v}(R)$
- B(R)= size of R in blocks
- T(R) = number of tuples in R
- V(R, a) = # of distinct values of attribute a

What is the cost in each case?
- Clustered index on a: B(R)/V(R,a)
- Unclustered index on a: T(R)/V(R,a)

Note: we ignore I/O cost for index pages

24

Index Based Selection

- Example: B(R) = 2000, T(R) = 100,000, V(R, a) = 20 — cost of $\sigma_{a=v}(R)$ = ?
- Table scan:
- Index based selection:

25

Index Based Selection

- Example: B(R) = 2000, T(R) = 100,000, V(R, a) = 20 — cost of $\sigma_{a=v}(R)$ = ?
- Table scan: B(R) = 2,000 I/Os
- Index based selection:

26

Index Based Selection

- Example: B(R) = 2000, T(R) = 100,000, V(R, a) = 20 — cost of $\sigma_{a=v}(R)$ = ?
- Table scan: B(R) = 2,000 I/Os
- Index based selection:
  - If index is clustered:
  - If index is unclustered:

27

Index Based Selection

- Example: B(R) = 2000, T(R) = 100,000, V(R, a) = 20 — cost of $\sigma_{a=v}(R)$ = ?
- Table scan: B(R) = 2,000 I/Os
- Index based selection:
  - If index is clustered: B(R)/V(R,a) = 100 I/Os
  - If index is unclustered:

28

Index Based Selection

- Example: B(R) = 2000, T(R) = 100,000, V(R, a) = 20 — cost of $\sigma_{a=v}(R)$ = ?
- Table scan: B(R) = 2,000 I/Os
- Index based selection:
  - If index is clustered: B(R)/V(R,a) = 100 I/Os
  - If index is unclustered: T(R)/V(R,a) = 5,000 I/Os

29

Index Based Selection

- Example:

B(R) = 2000
T(R) = 100,000
V(R, a) = 20

cost of $\sigma_{a=v}(R)$ = ?

- Table scan: B(R) = 2,000 I/Os!
- Index based selection:
  - If index is clustered: B(R)/V(R,a) = 100 I/Os
  - If index is unclustered: T(R)/V(R,a) = 5,000 I/Os!

30

Index Based Selection

- Example:

B(R) = 2000
T(R) = 100,000
V(R, a) = 20

cost of $\sigma_{a=v}(R)$ = ?

- Table scan: B(R) = 2,000 I/Os!
- Index based selection:
  - If index is clustered: B(R)/V(R,a) = 100 I/Os
  - If index is unclustered: T(R)/V(R,a) = 5,000 I/Os!

Lesson: Don't build unclustered indexes when V(R,a) is small !

31

Index Based Selection

- Example:

B(R) = 2000
T(R) = 100,000
V(R, a) = 20

cost of $\sigma_{a=v}(R)$ = ?

- Table scan: B(R) = 2,000 I/Os
- Index based selection:
  - If index is clustered: B(R)/V(R,a) = 100 I/Os
  - If index is unclustered: T(R)/V(R,a) = 5,000 I/Os

Lesson: Don't build unclustered indexes when V(R,a) is small !

32

Index Nested Loop Join

$R \bowtie S$

- Assume S has an index on the join attribute
- Iterate over R, for each tuple fetch corresponding tuple(s) from S
- Previous nested loop join: cost
  - B(R) + T(R)*B(S)
- Index Nested Loop Join Cost:
  - If index on S is clustered: B(R) + T(R)B(S)/V(S,a)
  - If index on S is unclustered: B(R) + T(R)T(S)/V(S,a)

33

Outline

- **Join operator algorithms**
  - One-pass algorithms (Sec. 15.2 and 15.3)
  - Index-based algorithms (Sec 15.6)
  - Two-pass algorithms (Sec 15.4 and 15.5)

34

Two-Pass Algorithms

- Fastest algorithm seen so far is one-pass hash join
  What if data does not fit in memory?
- Need to process it in multiple passes
- Two key techniques
  - Sorting
  - Hashing

35

Basic Terminology

- A run in a sequence is an increasing subsequence
- What are the runs?

2, 4, 99, 103, 88, 77, 3, 79, 100, 2, 50

36

Basic Terminology

- A run in a sequence is an increasing subsequence
- What are the runs?

2, 4, 99, 103,|88,|77,|3, 79, 100,|2, 50

37

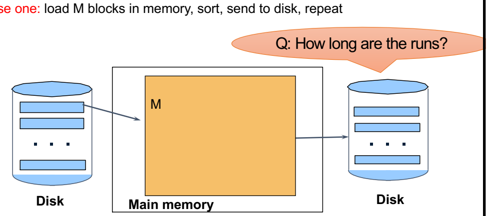
External Merge-Sort: Step 1

Phase one: load M blocks in memory, sort, send to disk, repeat

41

External Merge-Sort: Step 1

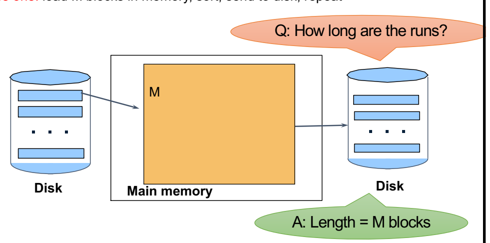
Phase one: load M blocks in memory, sort, send to disk, repeat



42

External Merge-Sort: Step 1

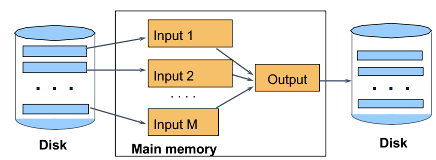
Phase one: load M blocks in memory, sort, send to disk, repeat



43

Phase two: merge M runs into a bigger run

- Merge M – 1 runs into a new run
- Result: runs of length M (M – 1) ≈ M^2



45

Example

- Merging three runs to produce a longer run:

0, 14, 33, 88, 92, 192, 322
2, 4, 7, 43, 78, 103, 523
1, 6, 9, 12, 33, 52, 88, 320

Output:
0

46

Example

- Merging three runs to produce a longer run:

0, **14, 33, 88, 92, 192, 322**
2, 4, 7, 43, 78, 103, 523
1, 6, 9, 12, 33, 52, 88, 320

Output:
0, ?

47

Example

- Merging three runs to produce a longer run:

0, **14, 33, 88, 92, 192, 322**
2, 4, 7, 43, 78, 103, 523
1, **6, 9, 12, 33, 52, 88, 320**

Output:
0, 1, ?

48

Example

- Merging three runs to produce a longer run:

0, **14, 33, 88, 92, 192, 322**
2, 4, 7, **43, 78, 103, 523**
1, 6, **9, 12, 33, 52, 88, 320**

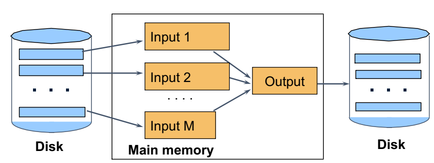
Output:
0, 1, 2, 4, 6, 7, ?

49

External Merge-Sort: Step 2

Phase two: merge M runs into a bigger run

- Merge M – 1 runs into a new run
- Result: runs of length M (M – 1) ≈ M^2



If approx. B <= M^2 then we are done

50

Cost of External Merge Sort

- Assumption: B(R) <= M^2
- Read+write+read = 3B(R)

51

Discussion

- What does $B(R) <= M^2$ mean?
- How large can R be?

52

Discussion

- What does $B(R) <= M^2$ mean?
- How large can R be?

- Example:
  - Page size = 32KB
  - Memory size 32GB: $M = 10^6$ pages

53

Discussion

- What does $B(R) <= M^2$ mean?
- How large can R be?

- Example:
  - Page size = 32KB
  - Memory size 32GB: $M = 10^6$ pages

- R can be as large as 10^{12} pages
  - 32×10^{15} Bytes = 32 PB

54

Merge-Join

Join $R \bowtie S$

- How?....

55

Merge-Join

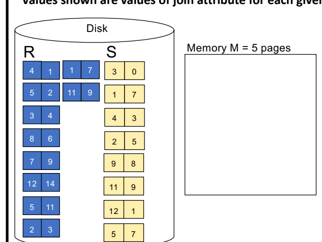
Join $R \bowtie S$

- Step 1a: generate initial runs for R
- Step 1b: generate initial runs for S
- Step 2: merge and join
  - Either merge first and then join
  - Or merge & join at the same time

56

Merge-Join Example

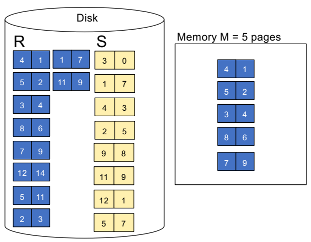
Setup: Want to join R and S
Relation R has 10 pages with 2 tuples per page
Relation S has 8 pages with 2 tuples per page
Values shown are values of join attribute for each given tuple



57

Merge-Join Example

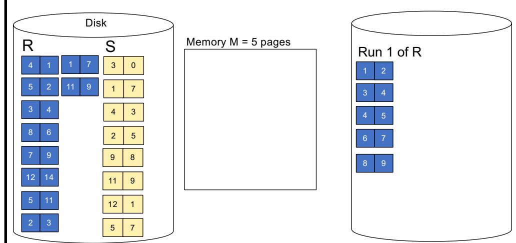
Step 1: Read M pages of R and sort in memory



58

Merge-Join Example

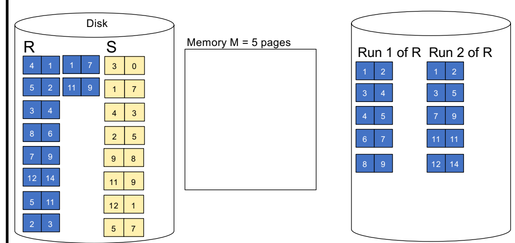
Step 1: Read M pages of R and sort in memory, then write to disk



59

Merge-Join Example

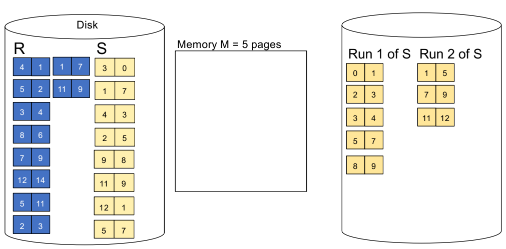
Step 1: Repeat for next M pages until all R is processed



60

Merge-Join Example

Step 1: Do the same with S

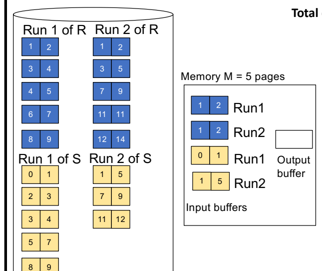


61

Merge-Join Example

Step 2: Join while merging sorted runs

Total cost: 3B(R) + 3B(S)



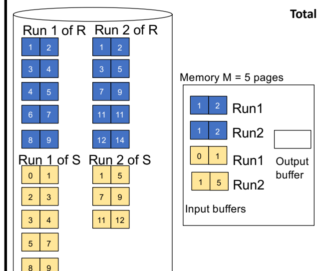
Step 2: Join while merging
Output tuples

62

Merge-Join Example

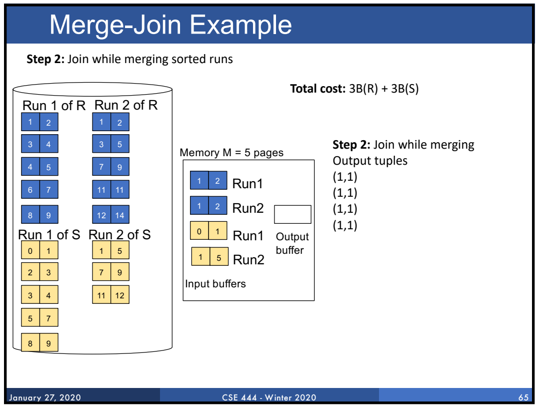
Step 2: Join while merging sorted runs

Total cost: 3B(R) + 3B(S)

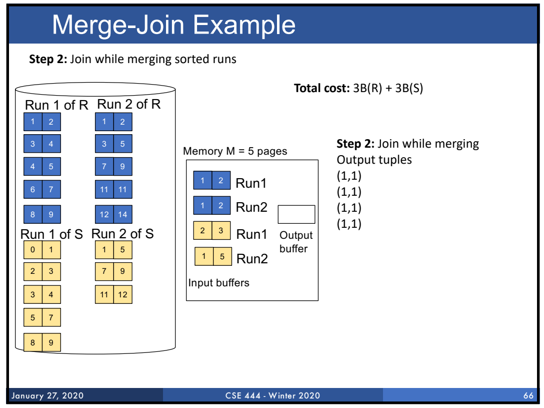


Step 2: Join while merging
Output tuples

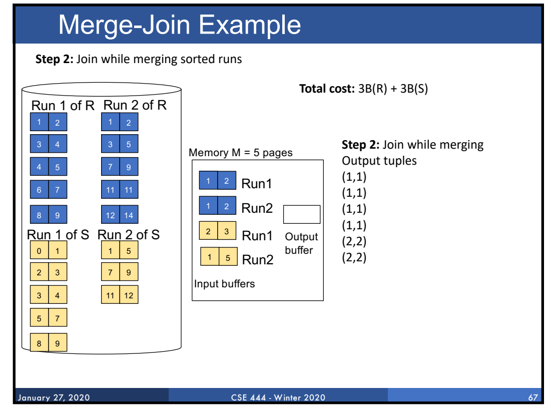
64



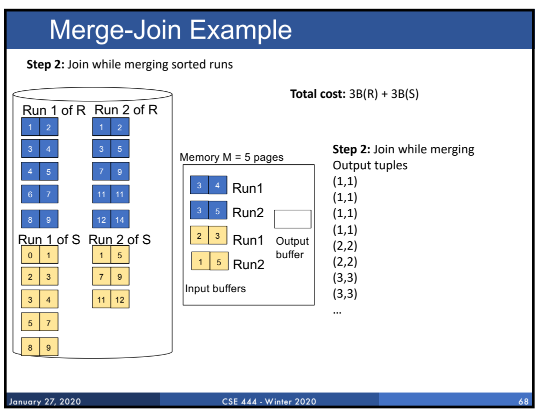
65



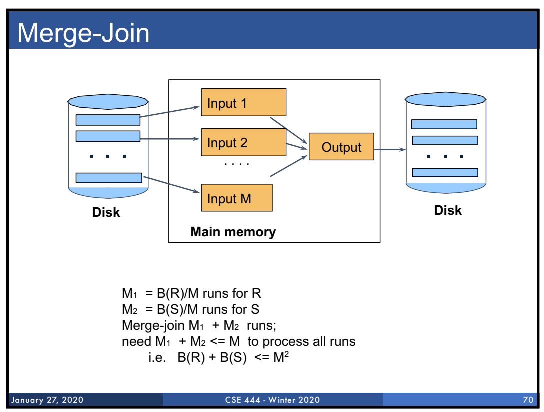
66



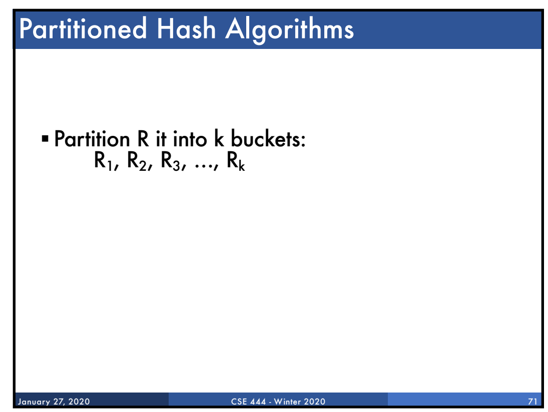
67



68



70



71

Partitioned Hash Algorithms

- Partition R it into k buckets:
  $R_1, R_2, R_3, \ldots, R_k$
- Assuming $B(R_1)=B(R_2)=\ldots= B(R_k)$, we have
  $B(R_i) = B(R)/k$, for all i

72

Partitioned Hash Algorithms

- Partition R it into k buckets:
  $R_1, R_2, R_3, \ldots, R_k$
- Assuming $B(R_1)=B(R_2)=\ldots= B(R_k)$, we have
  $B(R_i) = B(R)/k$, for all i
- Goal: each R_i should fit in main memory:
  $B(R_i) \le M$

73

Partitioned Hash Algorithms

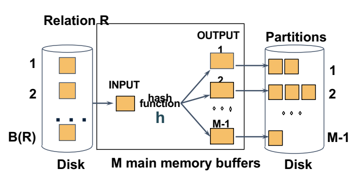
- Partition R it into k buckets:
  $R_1, R_2, R_3, \ldots, R_k$
- Assuming $B(R_1)=B(R_2)=\ldots= B(R_k)$, we have
  $B(R_i) = B(R)/k$, for all i
- Goal: each R_i should fit in main memory:
  $B(R_i) \le M$

How do we choose k?

74

Partitioned Hash Algorithms

- We choose k = M-1 Each bucket has size approx. B(R)/(M-1) ≈ B(R)/M

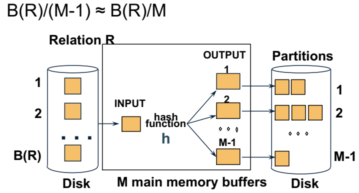


Assumption: $B(R)/M \le M$, i.e. $B(R) \le M^2$

75

Partitioned Hash Algorithms

- We choose k = M-1 Each bucket has size approx. B(R)/(M-1) ≈ B(R)/M



Assumption: $B(R)/M \le M$, i.e. $B(R) \le M^2$

76

Grace-Join

$R \bowtie S$

Note: grace-join is also called *partitioned hash-join*

77

Grace-Join

R ⋈ S

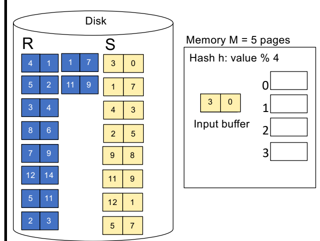
- Step 1:
  - Hash S into M-1 buckets
  - Send all buckets to disk
- Step 2
  - Hash R into M-1 buckets
  - Send all buckets to disk
- Step 3
  - Join every pair of buckets

Note: grace-join is also called *partitioned hash-join*

78

Partitioned Hash-Join Example

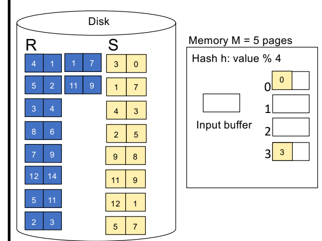
Step 1: Read relation S one page at a time and hash into M-1 (=4 buckets)



79

Partitioned Hash-Join Example

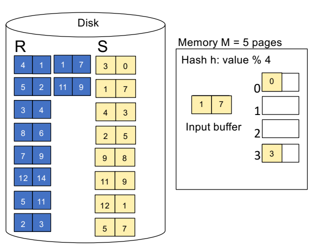
Step 1: Read relation S one page at a time and hash into the 4 buckets



80

Partitioned Hash-Join Example

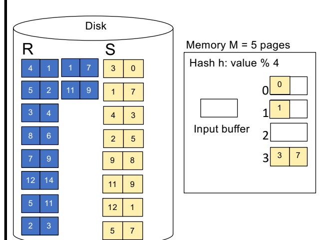
Step 1: Read relation S one page at a time and hash into the 4 buckets



81

Partitioned Hash-Join Example

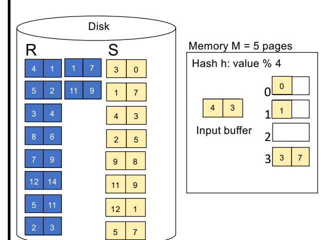
Step 1: Read relation S one page at a time and hash into the 4 buckets



82

Partitioned Hash-Join Example

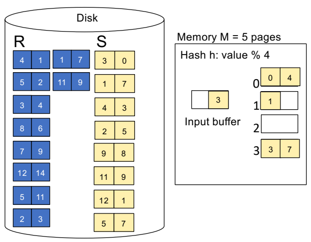
Step 1: Read relation S one page at a time and hash into the 4 buckets



83

Partitioned Hash-Join Example

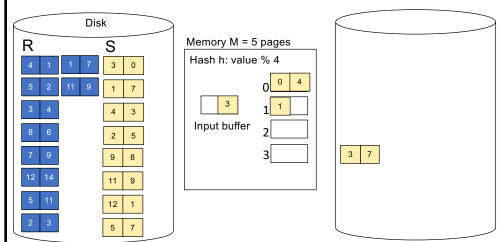
Step 1: Read relation S one page at a time and hash into the 4 buckets
When a bucket fills up, flush it to disk



84

Partitioned Hash-Join Example

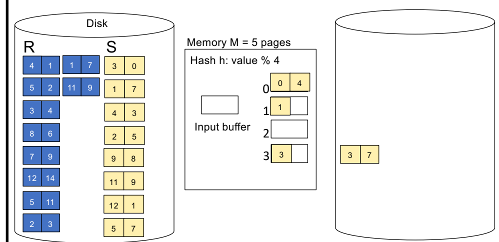
Step 1: Read relation S one page at a time and hash into the 4 buckets
When a bucket fills up, flush it to disk



85

Partitioned Hash-Join Example

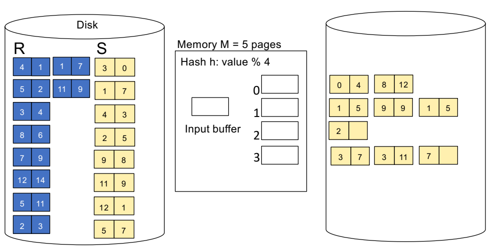
Step 1: Read relation S one page at a time and hash into the 4 buckets
When a bucket fills up, flush it to disk



86

Partitioned Hash-Join Example

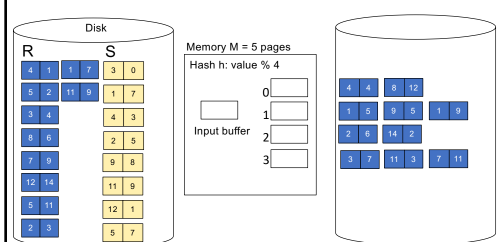
Step 1: Read relation S one page at a time and hash into the 4 buckets
At the end, we get relation S back on disk split into 4 buckets



87

Partitioned Hash-Join Example

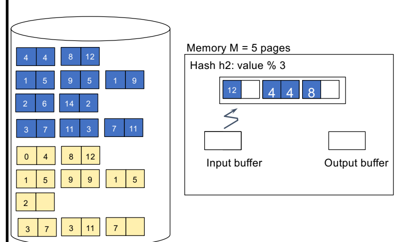
Step 2: Read relation R one page at a time and hash into same 4 buckets



88

Partitioned Hash-Join Example

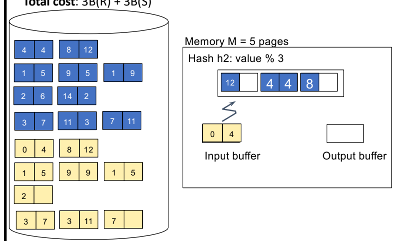
Step 3: Read one partition of R and create hash table in memory using a ***different*** hash function



89

Partitioned Hash-Join Example

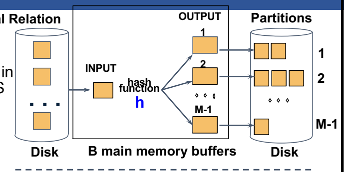
Step 4: Scan matching partition of S and probe the hash table
Step 5: Repeat for all the buckets
Total cost: 3B(R) + 3B(S)



90

Grace-Join

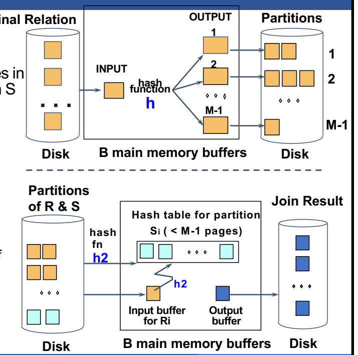
- Partition both relations using hash fn **h**: R tuples in partition i will only match S tuples in partition i.



91

Grace-Join

- Partition both relations using hash fn **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.



92

Grace Join

- Cost: 3B(R) + 3B(S)
- Assumption: min(B(R), B(S)) <= M^2

93

Hybrid Hash Join Algorithm

- Partition S into k buckets
  t buckets S_1 , …, S_t stay in memory
  k-t buckets S_{t+1}, …, S_k to disk
- Partition R into k buckets
  - First t buckets join immediately with S
  - Rest k-t buckets go to disk
- Finally, join k-t pairs of buckets:
  (R_{t+1},S_{t+1}), (R_{t+2},S_{t+2}), …, (R_k,S_k)

94

Summary of External Join Algorithms

- Block Nested Loop: B(S) + B(R)*B(S)/(M-1)
- Index Join: B(R) + T(R)B(S)/V(S,a)
  (unclustered)
- Partitioned Hash: 3B(R)+3B(S);
  - min(B(R),B(S)) <= M^2
- Merge Join: 3B(R)+3B(S)
  - B(R)+B(S) <= M^2

109

Summary of Query Execution

- For each logical query plan
  - There exist many physical query plans
  - Each plan has a different cost
  - Cost depends on the data
- Additionally, for each query
  - There exist several logical plans
- Next lecture: query optimization
  - How to compute the cost of a complete plan?
  - How to pick a good query plan for a query?

110