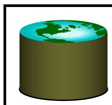
External Sorting

Chapter 13

One of the advantages of being disorderly is that one is constantly making exciting discoveries.

A. A. Milne

1



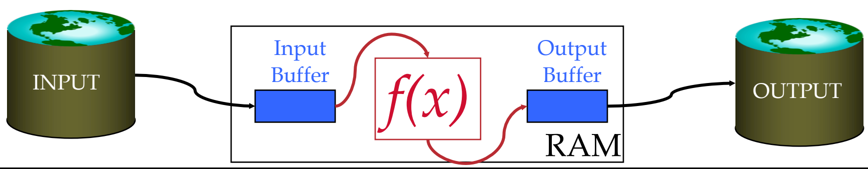
Why Sort?

- **A classic problem in computer science!**
- **Data requested in sorted order**
  - e.g., find students in increasing *gpa* order
- **Sorting is first step in *bulk loading* B+ tree index.**
- **Sorting useful for eliminating *duplicate copies* in a collection of records (Why?)**
- **Sorting is useful for summarizing related groups of tuples**
- ***Sort-merge* join algorithm involves sorting.**
- **Problem: sort 1Gb of data with 1Mb of RAM.**
  - why not virtual memory?

2

Streaming Data Through RAM

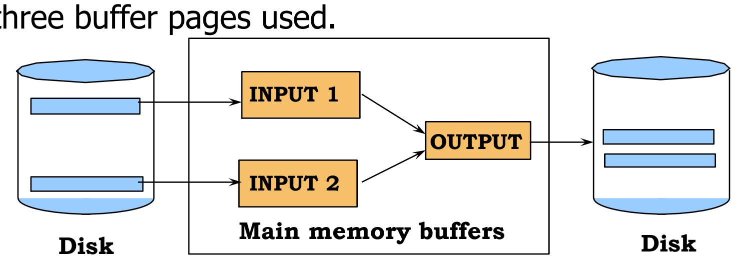
- **An important detail for sorting & other DB operations**
- **Simple case:**
  - Compute *f(x)* for each record, write out the result
  - Read a page from INPUT to Input Buffer
  - Write *f(x)* for each item to Output Buffer
  - When Input Buffer is consumed, read another page
  - When Output Buffer fills, write it to OUTPUT
- **Reads and Writes are *not* coordinated**
  - E.g., if f() is Compress(), you read many pages per write.
  - E.g., if f() is DeCompress(), you write many pages per read.



3

2-Way Sort

- **Pass 0: Read a page, sort it, write it.**
  - only one buffer page is used (as in previous slide)
- **Pass 1, 2, ..., etc.:**
  - requires 3 buffer pages
  - merge pairs of runs into runs twice as long
  - three buffer pages used.



4

Two-Way External Merge Sort

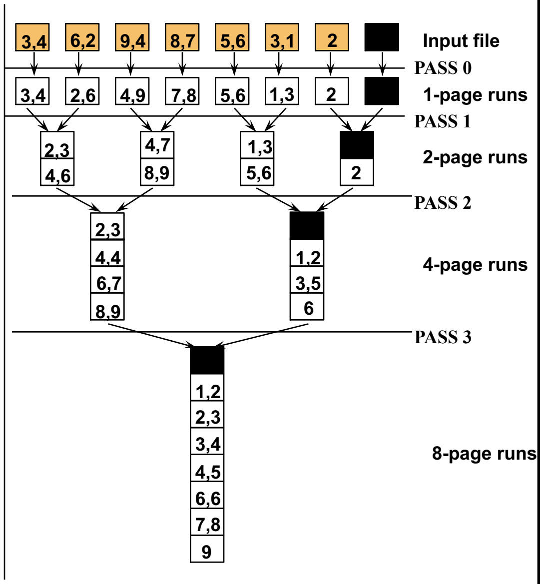
- **Each pass we read + write each page in file.**
- **N pages in the file => the number of passes**

$$= \lceil \log_2 N \rceil + 1$$

- **So total cost is:**

$$2N\left(\lceil \log_2 N \rceil + 1\right)$$

- ***Idea:*** *Divide and conquer:* **sort subfiles and merge**

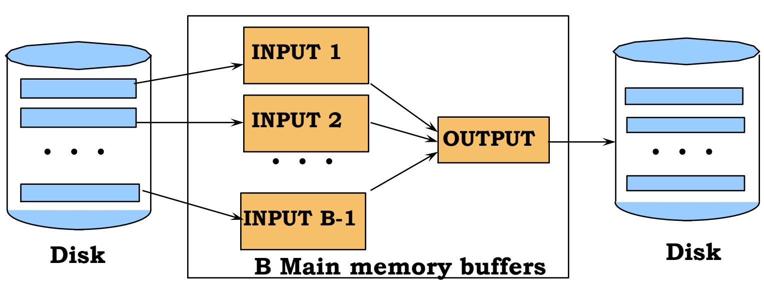


5

General External Merge Sort

☛ *More than 3 buffer pages. How can we utilize them?*

- **To sort a file with *N* pages using *B* buffer pages:**
  - Pass 0: use *B* buffer pages. Produce $\lceil N/B \rceil$ sorted runs of *B* pages each.
  - Pass 1, 2, …, etc.: merge *B-1* runs.



6

Cost of External Merge Sort

- **Number of passes:** $1+\lceil \log_{B-1}\lceil N/B\rceil\rceil$
- **Cost = 2N * (# of passes)**
- **E.g., with 5 buffer pages, to sort 108 page file:**
  - Pass 0: $\lceil 108/5\rceil$ = 22 sorted runs of 5 pages each (last run is only 3 pages)
- **Now, do four-way (B-1) merges**
  - Pass 1: $\lceil 22/4\rceil$ = 6 sorted runs of 20 pages each (last run is only 8 pages)
  - Pass 2: 2 sorted runs, 80 pages and 28 pages
  - Pass 3: Sorted file of 108 pages

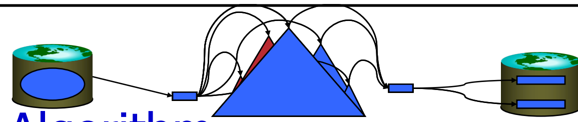
7

Number of Passes of External Sort

(I/O cost is 2N times number of passes)

N	B=3	B=5	B=9	B=17	B=129	B=257
100	7	4	3	2	1	1
1,000	10	5	4	3	2	2
10,000	13	7	5	4	2	2
100,000	17	9	6	5	3	3
1,000,000	20	10	7	5	3	3
10,000,000	23	12	8	6	4	3
100,000,000	26	14	9	7	4	4
1,000,000,000	30	15	10	8	5	4

8



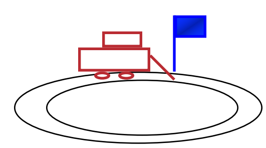
Internal Sort Algorithm

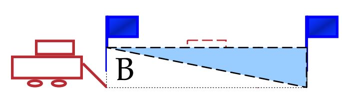
- **Quicksort is a fast way to sort in memory.**
- **Alternative: "tournament sort" (a.k.a. "heapsort", "replacement selection")**
- **Keep two heaps in memory, H1 and H2**

```
read B-2 pages of records, inserting into H1;
while (records left) {
    m = H1.removemin();  put m in output buffer;
    if (H1 is empty)
        H1 = H2;  H2.reset();  start new output run;
    else
        read in a new record r (use 1 buffer for
          input pages);
        if (r < m)  H2.insert(r);
        else        H1.insert(r);
}
H1.output();  start new run;  H2.output();
```


More on Heapsort

- **Fact: average length of a run is *2(B-2)***
  - The "snowplow" analogy
- **Worst-Case:**
  - What is min length of a run?
  - How does this arise?
- **Best-Case:**
  - What is max length of a run?
  - How does this arise?
- **Quicksort is faster, but ... longer runs often means fewer passes!**





I/O for External Merge Sort

- **Actually, doing I/O a page at a time**
  - Not an I/O per record
- **In fact, read a *block (chunk)* of pages sequentially!**
- **Suggests we should make each buffer (input/output) be a *chunk* of pages.**
  - But this will reduce fan-out during merge passes!
  - In practice, most files still sorted in 2-3 passes.

11

Number of Passes of Optimized Sort

N	B=1,000	B=5,000	B=10,000
100	1	1	1
1,000	1	1	1
10,000	2	2	1
100,000	3	2	2
1,000,000	3	2	2
10,000,000	4	3	3
100,000,000	5	3	3
1,000,000,000	5	4	3

☛ *Block size = 32, initial pass produces runs of size 2B.*

12

Sorting Records!

- **Sorting has become a blood sport!**
  - Parallel sorting is the name of the game ...
- **Minute Sort: how many 100-byte records can you sort in a minute?**
  - Typical DBMS: 10MB (~100,000 records)
  - Current World record: 55 TB
  - 394 Alibaba Cloud ECS ecs.n1.large nodes x
    (Haswell E5-2680 v3, 8 GB memory,
    40GB Ultra Cloud Disk, 4x 135GB SSD Cloud Disk)
- **Gray Sort: Sort rate (TBs / minute) achieved while sorting a very large amount of data (currently 100 TB minimum).**
  - 100 TB in 98.8 Seconds
    512 nodes x (2 OpenPOWER 10-core POWER8 2.926 GHz,
    512 GB memory, 4x Huawei ES3600P V3 1.2TB NVMe SSD,
    100Gb Mellanox ConnectX4-EN)
- **See http://sortbenchmark.org/**

13

Sorting Records!

- **CloudSort: Metric: Minimum cost for sorting a very large amount of data on a public cloud. (currently 100 TB).**
  - 100 TB for $144
    394 Alibaba Cloud ECS ecs.n1.large nodes x
    (Haswell E5-2680 v3, 8 GB memory,
    40GB Ultra Cloud Disk, 4x 135GB SSD Cloud Disk)
- **See http://sortbenchmark.org/**

14

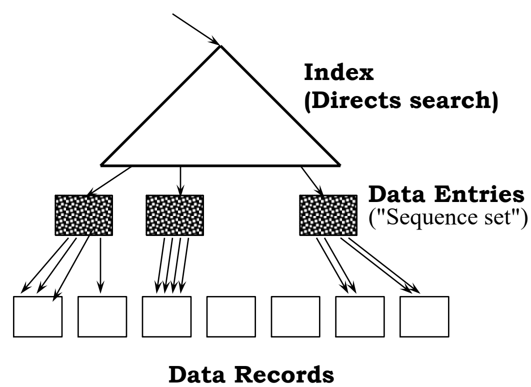
Using B+ Trees for Sorting

- **Scenario: Table to be sorted has B+ tree index on sorting column(s).**
- **Idea: Can retrieve records in order by traversing leaf pages.**
- *Is this a good idea?*
- **Cases to consider:**
  - B+ tree is clustered ***Good idea!***
  - B+ tree is not clustered ***Could be a very bad idea!***

15

Clustered B+ Tree Used for Sorting

- **Alternative 1**
- **Cost: root to the left-most leaf, then retrieve all leaf pages'**
- **Additional cost of retrieving data records: each page fetched just once.**

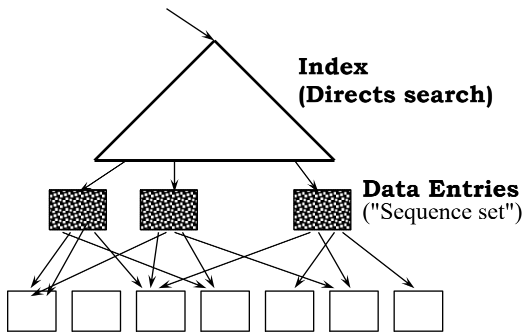


☛ *Better than external sorting!*

16

Unclustered B+ Tree Used for Sorting

- **Alternative (1) for data entries; each data entry contains *rid* of a data record. In general, one I/O per data record!**



17

External Sorting vs. Unclustered Index

N	Sorting	p=1	p=10	p=100
100	200	100	1,000	10,000
1,000	2,000	1,000	10,000	100,000
10,000	40,000	10,000	100,000	1,000,000
100,000	600,000	100,000	1,000,000	10,000,000
1,000,000	8,000,000	1,000,000	10,000,000	100,000,000
10,000,000	80,000,000	10,000,000	100,000,000	1,000,000,000

- *p*: # of records per page
- *B=1,000 and block size=32 for sorting*
- *p=100 is the more realistic value.*

18

Summary

- **External sorting is important; DBMS may dedicate part of buffer pool for sorting!**
- **External merge sort minimizes disk I/O cost:**
  - Pass 0: Produces sorted ***runs*** of size ***B*** (# buffer pages). Later passes: ***merge*** runs.
  - # of runs merged at a time depends on ***B**,* and ***block size****.*
  - Larger block size means less I/O cost per page.
  - Larger block size means smaller # runs merged.
  - In practice, # of passes rarely more than 2 or 3.

19

Summary, cont.

- **Choice of internal sort algorithm may matter:**
  - Quicksort: Quick!
  - Heap/tournament sort: slower (2x), longer runs
- **The best sorts are wildly fast:**
  - Despite 40+ years of research, we're still improving!
- **Clustered B+ tree is good for sorting; unclustered tree is usually very bad.**

20