CS 186 Spring 2024

Introduction to Database Systems
Lakshya Jain

DIS 5

1 General External Merge Sort

Given – N pages to sort, B buffer pages in memory

Pass 1 – Use B buffer pages. Produce $\frac{N}{B}$ sorted runs of B pages each.

Further passes – Merge $B-1$ runs.

Last pass – Produces 1 run of N pages.

Total I/O cost – $2N$ (# of passes) = $2N\left(1+\lceil log_{B-1}\lceil\frac{N}{B}\rceil\rceil\right)$

(a) You have 4 buffer pages and your file has a total of 108 pages of records to sort. How many passes would it take to sort the file?

(b) How many runs would each pass produce?

(c) What is the total cost for this sort process in terms of I/O?

(d) If the pages were already sorted individually, how many passes would it take to sort the file and how many I/Os would it be instead?

(e) If we wanted to sort N pages in at most p total passes, write an expression relating the minimum # of buffer pages B needed with N and p. What do you notice about B when $p = 1$?

2 Hashing

Given – *N* pages to hash, *B* buffer pages in memory
Initial partitioning – Read in *N* pages and hash into *B - 1* partitions. Write out

$$\sum_{i=1}^{B-1} (\text{\# of pages in partition i})$$

Recursive partitioning – For each individual partition, recursively partition if its size $s > B$.
Building in-memory hash table – Once a partition's size $s \leq B$, read in *s* pages, build an in-memory hash table, and write out *s* pages.

(a) What are some use-cases in which hashing is preferred over sorting?

(b) Suppose we have *B* buffer pages and can process *B(B - 1)* pages of data with External Hashing in two passes. For this case, fill in the blanks with the appropriate # of pages.

_____ input buffer(s)

_____ partitions after pass 1

_____ pages per partition

(c) If you are processing exactly *B(B - 1)* pages of data with external hashing, is it likely that you'll have to perform recursive external hashing? Why or why not?

(d) If we have 10 buffer pages, what is the maximum number of pages we could externally hash in 3 passes? Assume a perfectly uniform hash function for each pass.

(e) We want to hash *N = 100* pages using *B = 10* buffer pages. Suppose in the initial partitioning pass, the pages are unevenly hashed into partitions of 10, 20, 20, and 50 pages. Assuming uniform hash functions are used for every partitioning pass after this pass, what is the total I/O cost for External Hashing?