COMP 3005: Database Management Systems **(Due: N/A)**

Assignment #6

Instructor: Ahmed El-Roby *Name: , ID:*

Instructions: Read all the instructions below carefully before you start working on the assignment, and before you make a submission.

- The accepted formats for your submission are: pdf and docx. More details below.
- If you use the tex file, make sure you edit line 28 to add your name and ID. Only write your solution and do not change anything else in the tex file. If you do, you will be penalized.
- No late submissions are allowed.

Q 1: (4 points)

Use the external sort-merge algorithm to sort the following relation:
(k), (w), (e), (y), (p), (l), (x), (z), (m), (i), (h), (b).
Assume that only one tuple fits in a block and memory holds at most three blocks for sorting. Show your steps.
Sorted runs: (e, k, w), (l, p, y), (m, x, z), and (b, h, i).
First Merge Pass:
- Merging 1st and 2nd runs: (e, k, l, p, w, y).
- Merging 3rd and 4th runs: (b, h, i, m, x).
Second Merge Pass:
- Merging 1st and 2nd runs: (b, e, h, k, l, m, p, w, x, y, z).

Q 2: (4 points)

During the merge pass(es), reading in each run one block at a time leads to a large number of seeks. To reduce the number of seeks, a larger number of blocks, denoted b_b are read or written at a time. This means that b_b buffer blocks are allocated to each input run and to output run. What is the total number of block transfers for external sorting of a relation? What is the total number of seeks? Explain your answer.
Hint: Using $b_b > 1$ could increase the number of merge passes. Thus, more block transfers. But less number of seeks.
Block transfers: $b_r(2\lceil log_{\lfloor M/b_b \rfloor - 1}(b_r/M)\rceil + 1)$.
Seeks: $2\lceil b_r/M\rceil + \lceil b_r/b_b\rceil (2\lceil log_{\lfloor M/b_b \rfloor - 1}(b_r/M)\rceil - 1)$.

Q 3: (10 points)

Suppose you need to sort a relation of 40 gigabytes, with 4-kilobyte blocks, using a memory size of 40 megabytes. Suppose the cost of a seek is 5 milliseconds, while the disk transfer rate is 40 megabytes per second. Assume $b_b = 1$.
Now, assume a flash storage device is used instead of a disk, and it has a latency of 20 microsecond and a transfer rate of 400 megabytes per second. Recompute the cost of sorting the relation, in seconds.
As an exercise (no marks), resolve the problem with $b_b = 100$ blocks.
$M = \frac{40\times10^6}{4\times10^3} = 10^4$.
$b_r = \frac{40\times10^9}{4\times10^3} = 10^7$.
Disk seek cost is 5×10^{-3} seconds.
Block transfer time is 10^{-4} seconds ($\frac{4\times10^3}{40\times10^6}$).
Total cost is (Number of disk seeks $\times$ Disk seek cost) + (Number of block transfers $\times$ Block transfer time).
The number of block transfers is: $10^7(2 \times \left\lceil log_{10^4-1}\frac{10^7}{10^4}\right\rceil + 1) = 3 \times 10^7$ block transfers.
Time spent in block transfers is: $10^{-4} \times 3 \times 10^7 = 3000$ seconds.
The number of seeks is: $2 \times \frac{10^7}{10^4} + 10^7(2 \times \left\lceil log_{10^4-1}\frac{10^7}{10^4}\right\rceil - 1) = 2000 + 10^7$ seeks.

Time spent in seeks is: $5 \times 10^{-3} \times (2000 + 10^7) = 50010$ seconds.
Total time is: $3000 + 50010 = 53010$ seconds.

For flash storage, latency is 20 microseconds, and block transfer time is $\frac{4 \times 10^3}{400 \times 10^6} = 10^{-5}$ seconds.
Time spent in block transfers is: $10^{-5} \times 3 \times 10^7 = 300$ seconds.
Time spent in read latencies is: $20 \times 10^{-6} \times (2000 + 10^7) = 200.04$ seconds.
Total time is: $300 + 200.04 = 500.04$ seconds.