The University of British Columbia

Computer Science 404: Database Systems

Midterm Examination #1, February 9, 2005. Instructor: E.M. Knorr

Time: 48 minutes. Only a simple, non-programmable calculator is permitted.
Closed book. No notes.

Name_____ Student No_____

(PRINT) (Last) (First)

Signature_____

The examination has 5 pages, and that includes this cover sheet. Check that you have a complete paper.

Print your name and ID at the top of this page, and provide your signature. Have your student ID ready.

In the interest of time, it is sufficient to print your initials at the top of pages 2-5.

A simple, non-programmable, non-communicating calculator is permitted. No other aids are allowed.

Work quickly and do the easy questions first. Part marks are available.

The marks for each question are given in braces. **Do not spend too much time on any one question.**

To minimize disruptions during the exam, **please avoid asking the invigilators for help, tips, or explanations.** *Please write down any reasonable assumptions that you are making, if you believe that a question is ambiguous.*

Marks

Page	Max.	Achieved
2	7	
3	8	
4	8	
5	7	
Total	30	

1. {2 marks} Under what circumstance would it be wise to use a fixed length field to hold some text, instead of using a variable length field? Give one good reason.

2. {2 marks} Does mirroring on a DBMS improve write performance? Give a brief (one sentence) justification for your answer.

3. {3 marks} Consider the following RAID-6 Hamming matrix given in class, for 4 data disks and 3 redundant disks:

```
1 1 1 0 | 1 0 0
1 1 0 1 | 0 1 0
1 0 1 1 | 0 0 1
```

Suppose the data disks are:

Disk 1: `????????`
Disk 2: `11110000`
Disk 3: `????????`
Disk 4: `11111111`

and the redundant disks are:

Disk 5: `11001100`
Disk 6: `00000000`
Disk 7: `11000011`

Suppose disks 1 and 3 fail. Recover their contents.

2. {8 marks} Consider the following disk geometry:

Seek time = 0 if on the same cylinder, else (1 ms setup time + 1 ms for every 300 cylinders moved)
Rotation speed = 5400 rpm (revolutions per minute)
Transfer time = only the time it takes for the complete block to pass under the head
No parallel reads are allowed
Tracks per cylinder = 15
Number of cylinders = 4500
Sector size = 512 bytes
Sectors on a track = 256
Block (Page) size = 4K

(a) Estimate the *average* amount of time that it would take to read an entire cylinder of data, if the cylinder is to be randomly selected on the disk drive. Show your work. You can assume that we start reading the first track of a given cylinder immediately upon reaching it after a seek operation, without waiting for the starting block to come under the head.

(b) How many sequential/contiguous cylinders are required to hold 2000 pages?

3. {3 marks} We have a large table against which all kinds of queries are being asked—and some of these queries require a full scan of the table. Suppose Query X scans the table, and that it takes 10 minutes of elapsed time to answer the query. Someone suggests compressing the table in memory before writing it to disk, and uncompressing it after reading it in. Is it reasonable to expect Query X to take more time, less time, or about the same time to run the query on the compressed version? Explain briefly, using 1 or 2 sentences. (You can assume that the data can be compressed by 30-50%.)

4. {5 marks} Consider a dataset with 65,000,000 records, where 325 records fit on a 4K page. Furthermore, suppose that you have 400 4K buffer pages available.

Use a general multiway external mergesort, and show, for all the passes: the *number* of sorted runs (sublists) produced at each pass, and the *size* of the sorted runs (in pages) at each pass. Show your work.

5. {7 marks} Suppose we have 204,800 4K pages of data on contiguous cylinders, 32 MB of buffer space, and that we have 32 1-MB buffers available for use (where 1 cylinder = 1 MB). If a random seek takes 10 ms, and other seeks take (1 ms setup time + 1 ms/300 cylinders moved), and if a 4K page can be transferred in 0.5 ms, then estimate the amount of time (including a calculation for a reasonable amount of seek time, etc., even though it's likely to be small) for performing a *2-phase multiway merge sort using cylindrification*. (You can ignore *rotational latency* (only) when finding the "starting point" on a given cylinder.) Show your work for both phases.