CS-D64 Midterm Examination Name:_____

The legal stuff:

- The exam is open book and notes.
- Absolutely no communication with other students during the exam.

The numbers:

- You have 60 minutes to complete the exam.
- The maximum score is 60 points.

Steps to take:

1. Write your name at the top of each page.
2. Read over all the questions and ask for clarification early.
3. Note how much each question is worth and allocate your time proportionately.
4. Try to write your answer on the same page as the question - only if you absolutely cannot should you use scratch paper (put your name at the top of this as well).

Partial credit will be given where possible.

Relax. Write clearly. Good luck.

Question	Total	Score
1	5	
2	10	
3	5	
4	10	
5	15	
6	15	
	Total Score	

1) [5 points] True/False

For each of the following statements, circle T if the statement is true and F if the statement is false.

a. Usually, the time to read a block is dominated by the transfer time. (T/F)

b. In the best case, modifying a block requires the best case read time plus two full disk rotations. (T/F)

c. For a fixed size memory, a contiguous file with larger blocks will tend to require longer to sort than a contiguous file with smaller blocks. (T/F)

d. Even if a sector's checksum is correct, the sector could still have an error. (T/F)

e. It is possible in linear hashing that inserting an element can decrease the average search time of the index. (T/F)

2) [10 points] Mirroring versus striping

Disk striping is a technique in which two separate drives of capacity *c* are made to look like one drive of capacity 2c. Blocks 0,2,4,… of the virtual drive are actually blocks 0,1,2,… of the first small drive, and blocks 1,3,5,… of the virtual drive are blocks 0,1,2,… of the other small drive. What is the benefit of this technique? Compare it to mirroring for the following operations (be succinct):

A. Time to read a block

B. Time to read a large contiguous file

C. Time to write a block

D. Time to write a large contiguous file

Are there are any other relevant differences between the two techniques?

3) [5 points] Sorting

There's a well known tweak to the first phase of our sorting routine that allows sorted sublists of 2M blocks to be written out to disk rather than sublists of only M blocks. Without worrying about the details of the improvement, what will its impact be on the sorting algorithm overall? In particular, how large a file can be sort in only two passes?

4) [10 points] Sorted files

Suppose we want to keep a file composed of variable length records in sorted order. Describe the block layout you would use as well as a short description of what would happen when a record is inserted, deleted, or modified.

5) [15 points] Extensible Hashing

[10 points] Part A: Without storing any additional information, would it be possible to drop the information in the nub of the blocks in the extensible hashing scheme in the book? Specifically, I'm referring to the number of bits used to determine membership in the block. If so, what changes to the algorithms would be required? What would the impact be? If not, give an example which shows how the scheme would break.

5) Extensible Hashing (cont.)

[5 points] Part B: With the data structures and algorithms described in the book, what problems will duplicate search values cause?

6) [15 points] Range queries

How many I/Os will it take to answer the following query assuming we use a B+ tree on gpa?

```
SELECT *
FROM Student
WHERE gpa>3.8
```

You may use the following assumptions:

1. The student relation contains 40,000 tuples, each of which is 400 bytes.
2. Blocks are 4kb.
3. The students' GPAs, which are 8 byte floats, are uniformly distributed over the range [2.0,4.0].
4. Database pointers and block pointers are 5 bytes.
5. The index nodes are approximately 80% occupied.

Part A: How many I/Os will be required if the B-Tree is a primary index?

6) Range queries (cont.)

Part B: How many I/Os will be required if the B-Tree is a secondary index?