Database Systems

Instructor: Hao-Hua Chu

Fall Semester, 2005

Assignment 3: Storage and Indexing

Deadline: At the end of TA's office hour on Tuesday, Nov. 8, 2005
This is an individual assignment, that is, no group submissions are allowed.

Cheating Policy: If you are caught cheating, your grade is 0.
Late Policy: You may hand in your late assignment at TA's office hour on Wed. (11/9/2005) for 80% of original grade, or at TA's office hour on Thurs. (11/10/2005) for 70%. We will not accept any assignment submissions after Thursday.

Questions

1. Consider the following relations:
   Emp(*eid:* integer, *ename:* varchar, *sal:* integer, *age:* integer, *did:* integer)
   Dept(*did:* integer, *budget:* integer, *floor:* integer, *mgr_eid:* integer)
   Salaries range from $10,000 to $100,000, ages vary from 20 to 80, each department has about five employees on average, there are 10 floors, and budgets vary from $10,000 to $1 million.
   For each of the following queries, which of the listed index choices would you choose to speed up the query? Please also compute the number of page I/Os for each indexing choices: There are 100 data pages when records are packed onto pages with no wasted space and 100 records per page. The fan-out of B+ tree is 100. As the assumption in the textbook, each data entry in the index is one tenth the size of a data record. In B+ tree files, the pages are at 67% occupancy; in hash files, the pages are at 80% occupancy. Note that clustered implies Alternative 1 (page 17 in 10/31 slide).
   If your database system does not consider index-only plans (i.e., data records are always retrieved even if enough information is available in the index entry), how would your answer change? Explain briefly.
   (1) Query: Print ename, age, and sal for all employees.
      (a) Clustered hash index on <ename, age, sal> fields of Emp.
      (b) Unclustered hash index on <ename, age, sal> fields of Emp.
      (c) Clustered B+ tree index on <ename, age, sal> fields of Emp.

(d) Unclustered hash index on <eid, did> fields of Emp.

(e) No index.

(2)Query: Find the dids of departments that are on the 10^{th} floor and have a budget of less than $15,000. (Note: we assume applying the hash function to a record allows us to identify and retrieve the page containing the record with 1 I/O and that there are no overflow chains. We also assume that the matched records are all in the same page.)

(a) Clustered hash index on the floor field of Dept.

(b) Unclustered hash index on the floor field of Dept.

(c) Clustered B+ tree index on <floor, budget> fields of Dept.

(d) Clustered B+ tree index on the budget field of Eept.

(e) No index.

2. Consider the page format for variable-length records that uses a slot directory.
   (1) One approach to managing the slot directory is to use a maximum size (i.e., a maximum number of slots) and allocate the directory array when the page is created. Discuss the pros and cons of this approach with respect to the approach discussed in the text.
   (2) Suggest a modification to this page format that would allow us to sort records (according to the value in some field) without moving records and without changing the record ids.
3. Modern disk drives store more sectors on the outer tracks than the inner tracks. Since the rotation speed is constant, the sequential data transfer rate is also higher on the outer tracks. The seek time and rotational delay are unchanged. Given this information, explain good strategies for placing files with the following kinds of access patterns:
   (1) Frequent, random accesses to a small file (e.g., catalog relations).
   (2) Sequential scans of a large file (e.g., selection from a relation with no index).
   (3) Random accesses to a large file via an index (e.g., selection from a relation via the index).
   (4) Sequential scans of a small file.

Submission

Hand in PAPER PRINTOUT that contains your answers to the three questions. Please include your name and ID.