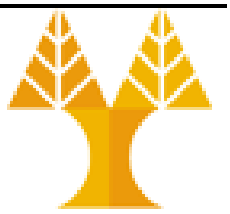
ΕΠΛ646 Προχωρημένα Θέματα
Βάσεων Δεδομένων



Τμήμα Πληροφορικής
Πανεπιστήμιο Κύπρου

ΕΡΓΑΣΤΗΡΙΟ #5

External Sorting and Evaluation of Relational Operators

1. **(Exercise 13.1)** Suppose you have a file with 10,000 pages and you have three buffer pages. Answer the following questions for each of these scenarios, assuming that our most general external sorting algorithm is used:
   i. A file with 10,000 pages and three available buffer pages.
   ii. A file with 20,000 pages and five available buffer pages.
   iii. A file with 2,000,000 pages and 17 available buffer pages.

   Questions
   i. How many runs will you produce in the first pass?
   ii. How many passes will it take to sort the file completely?
   iii. What is the total I/O cost of sorting the file?
   iv. How many buffer pages do you need to sort the file completely in just two passes?

2. **(Exercise 14.3)** Consider processing the following SQL projection query:

   SELECT DISTINCT E.title, E.ename FROM Executives E

   You are given the following information:
   Executives has attributes ename, title, dname, and address; all are string fields of the same length.
   The ename attribute is a candidate key.
   The relation contains 10,000 pages.
   There are 10 buffer pages.

   Consider the optimized version of the sorting-based projection algorithm: The initial sorting pass reads the input relation and creates sorted runs of tuples containing only attributes ename and title. Subsequent merging passes eliminate duplicates while merging the initial runs to obtain a single sorted result (as opposed to doing a separate pass to eliminate duplicates from a sorted result containing duplicates).

   i. How many sorted runs are produced in the first pass? What is the average length of these runs? (Assume that memory is utilized well and any available optimization to increase run size is used.) What is the I/O cost of this sorting pass?
   ii. How many additional merge passes are required to compute the final result of the projection query? What is the I/O cost of these additional passes?
   iii.
      a. Suppose that a clustered B+ tree index on *title* is available. Is this index likely to offer a cheaper alternative to sorting? Would your answer change if the index were unclustered? Would your answer change if the index were a hash index?
      b. Suppose that a clustered B+ tree index on *ename* is available. Is this index likely to offer a cheaper alternative to sorting? Would your answer change if the index were unclustered? Would your answer change if the index were a hash index?
      c. Suppose that a clustered B+ tree index on <*ename, title*> is available. Is this index likely to offer a cheaper alternative to sorting? Would your answer change if the index were unclustered? Would your answer change if the index were a hash index?
   iv. Suppose that the query is as follows:
      SELECT E.title, E.ename FROM Executives E
      That is, you are not required to do duplicate elimination. How would your answers to the previous questions change?