Database Tuning, ITU, Spring 2008

Rasmus Pagh

February 12, 2008

1 Choosing an index

In this problem we will investigate whether or not it can be beneficial to use two separate indexes on the same attribute for a relation, in a specific setting. The indexes are defined on a key for the relation, which has N records. We use B to denote the number of keys that fit in a disk block. We will consider two scenarios:

Scenario 1 Primary hash index on the key.

Scenario 2 Primary hash index **and** secondary B-tree index on the key.

Note that in both Scenario 1 and 2 the relation is only stored once, clustered according to the hash index. In the following we will assume for simplicity that the hash function used distributes the keys evenly, and that no overflow blocks are necessary, even after doing new insertions. Also, assume that the free space in B-tree nodes is negligible and can be ignored.

a) Suppose a key is 8 bytes and that a pointer is 4 bytes. What are the storage requirements of Scenario 1 and 2, besides the space used for storing the relation itself? (In giving the answer, you may ignore negligible terms.)

The following questions consider insertions of new records, point queries (to look up the record with a specific value for the key), and range queries (returning all records with key in a range of the form $[j, \ldots, j+r-1]$). We denote the size of the range by r, and the number of records returned by a range query by l.

b) Argue that it is possible to obtain the following I/O complexities for the operations insert, point query, and range query, in the two scenarios:

	Insert	**Point query**	**Range query**
Scenario 1	$O(1)$	$O(1)$	$O(r)$
Scenario 2	$O(\log_B N)$	$O(1)$	$O(\log_B N + l)$

In the following, ignore constant factors by assuming that all I/O bounds in **b)** hold with the constant 1 (i.e., ignore the big-O notation).

c) Consider a sequence of X insertions, Y point queries, and Z range queries, where range queries have range of average size r, and l records in the answer on average. When is scenario 1 and 2, respectively, the most efficient? Express your answer using an inequality involving (a subset of) the variables X, Y, Z, r, l, N, and B.