p5q1
MR

3 Data structures and algorithms 2004

It is proposed to store a large number of records on a disk using Larsen's method so that any lookup can be done using only one disk transfer. All the records are of length 200 bytes and each contains a 20 byte key. The data is to be held on a single disk preformated to contain 100,000,000 sectors each of size 4096 bytes. Reading multiple consecutive sectors is regarded as a single transfer.

(a) Describe Larsen's algorithm in detail and, for the records and disk specified above, state the disk block size, the signature size and the amount of main memory that you would choose to use. [10 marks]

(b) Carefully estimate the maximum number of records that could reasonably be stored on the disk assuming the sizes you gave in (a). [6 marks]

(c) Discuss the advantages and disadvantages of different signature sizes. [4 marks]

ANSWER NOTES:

(a) Description is bookwork, but the selection of sizes is not.

A 4096 byte sector can hold 20 x (200 byte records + signature value), it would probably be better to make the basic disk block larger, say 10 sector is 40960 bytes this could hold 204 records. So the hash chains can have a length of up to about 204 entries. There will be 10,000,000 blocks of this size so we need that number of entries in main memory for the signatures. A signature size of 8 bits is probably appropriate. These can be accessed efficiently and a storage requirement of 10,000,000 bytes is reasonable.

Think up two hash function blockhash(key, n) and sighash(key, n) The first to give block numbers in range 1 to 10,000,000 and the second to give signature values in the range 0 to 255. n = 0, 1, 2,... provides the sequence of block and signature hashes required by Larsen's method. For a record to be allowed in block b its signature hash must be less than or equal to the corresponding entry in the memory table. If there is insufficient free space in a block the corresponding main memory signature entry is reduces causing one or more current entries to be removed from the block and repositened elsewhere.

(b) With an 8-bit signature hash chain length can be controlled very well so it should be possible to run efficientely even when the blocks and on average 90 to 95% full. So the number of records that could reasonable be stored is about

10,000,000 * 204 * 0.95 = about 1,900,000,000

(c) The few bits in the signature the less control over the length of hash chains. Potentially this will make it harder to balance the hash chain lengths slightly reducing the practical number of records that can be strored. It does however save main memory. Some sizes such as bytes and 16 bit words are cheaper to access.