ENGINEERING TRIPOS PART IIA

Monday, 28 April 2003 2.30 – 4.00

Module 3I1

DATA STRUCTURES AND ALGORITHMS

*Answer **all** of Section A and **two** questions from Section B.*

All questions carry the same number of marks.

*The **approximate** percentage of marks allocated to each part of a question is indicated in the right margin.*

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator

(TURN OVER

SECTION A

Attempt all parts of this question

1 You should answer each of the following parts with a short-answer, quoting relevant results rather than proving them. Give names for algorithms you need to refer to and sketch methods or justifications of claims you make. Credit will be given for the clarity and succinctness of answers as well as for basic factual accuracy.

(a) If $f(n) = O(g(n))$ then is it true that $f(n) = \Theta(g(n))$? Similarly if $g(n)=\Theta(f(n))$ is it true that $g(n)=O(f(n))$ or $f(n)=O(g(n))$ or neither? [10%]

(b) Consider the function $push(n,m) = 2^n(2m+1)$. Explain how an implementation of the abstract datatype STACK could be built on the basis of this, in particular indicating the value of m to be used to denote an empty stack and explaining how the POP operation could be implemented. [10%]

(c) If $f(n) = 3f(n/3) + 7n$, what big-theta or big-O expressions can be used to characterize f(n)? [10%]

(d) Suppose that because of a programming error your hash function always returns the value 1. In terms of whatever parameters are relevant estimate the cost of inserting a new item into an already partly-filled hash table.. [10%]

(e) Describe how Larsen's method of dynamic hashing can be used to fetch a record from disc using just one disc block transfer, at the cost of holding a table in memory that stores an integer for each disc block. You should only describe how to retrieve data: no discussion of how to add new records is required here. [10%]

(f) Explain one scenario where a buddy system for storage allocation would out-perform first-fit, and another pattern of allocation and release where it would not. [10%]

(g) Give an upper and a lower bound for the number of nodes in a 2-3-4 tree of height h. Hence if a 2-3-4 tree holds n values find an upper bound on its height. [10%]

(Cont..

(h) Experience with the "zip" compression program suggests that the contents of very many computer files can be compressed to less than half their original size using Lempel Ziv. Is it reasonable to expect that by using Lempel Ziv twice they can be compressed further? [10%]

(i) What is a strongly-connected component of a graph G? [10%]

(j) What is a "winding number" and how can you use one to test if a point is inside or outside a given polygon? [10%]

SECTION B

*Answer **two** questions from this section.*

2 (a) One method of finding the median of a set of values starts by sorting them. Binary insertion sort is known to keep the number of comparisons that are performed low, even though it may involve excessive data movement. Explain how this sorting method works by showing what sequence of comparisons it could involve when sorting exactly five items. Work through examples so that you can show both the best-case and the worst-case number of comparisons involved in finding the median of five distinct values this way. [20%]

(b) Another median finding algorithm involves selecting a pivot and partitioning the data. For the case of just five values (again all known to be distinct) show what the best and worst-case number of comparisons are. [20%]

(c) A standard guaranteed linear-cost median finding algorithm organises its input data into groups of five and continues work with the median values from each of these sets of five. Establish a recurrence formula whose solution will confirm that the cost of this method is linear. [30%]

(d) Suppose now that instead of selecting a pivot by using the median of medians-of-five the algorithm is adjusted to form groups of seven. Find a recurrence formula that predicts the costs of the new method. Is its cost still linear? [15%]

(e) In a similar style one might try to simplify by forming groups of three rather than five. How do the expected costs grow with n in this case? [15%]

3 The SET datatype can be represented by keeping the items that are members of a set in a linked list. In such a list there will be no repeated items, but the order in which values appear can be arbitrary, and in particular at this stage you should consider the case where you cannot assume that the list has been sorted.

(a) For sets with n elements, represented this way, what are the expected costs of
  (i) Checking if an item is in the set; [10%]
  (ii) Adding a new item, given the knowledge that it is not already in the set; [10%]
  (iii) Forming the intersection and union of two sets using direct algorithms based on the above two operations. [20%]

(b) Devise intersection and union algorithms that start by inserting all items from one set into a hash table and continue by looking up all members of the second set in the same table. Assuming that each hash table operation has cost O(1) what overall cost predictions can you make? The hash table will be empty at the start of the procedure and you must leave it empty at the end. Your cost estimates must include allowing for this and for forming the result as a linked list. [25%]

(c) Alternatively suppose that the items in the list are objects that have a spare bit somewhere in them. As far as set operations are concerned items are only equal if they are actually represented as the same object. Devise intersection and union algorithms that start by scanning the first set and setting the spare bit in each object present in it, then scanning the second list checking which items have their mark bit set. All the mark bits will be zero to start with and you must leave them re-set to zero at the end of processing. Again give cost estimates. [25%]

(d) Compare the merits of the schemes, identifying any circumstances where you might want or need to use one rather than the other. [10%]

4 Suppose that a statistical model for a stream of characters indicates that only the digits 0 to 9 will appear, that each digit is equally probable and that there is no need to allow for an end-of-data marker. Then the idea behind arithmetic coding suggests that a string such as "752904..." would be associated with the number z=0.752904... and would be transmitted as a sequence of bits that were the binary representation of z.

(a) What is the binary representation for the value (1/3) that has decimal representation 0.33333...? [10%]

(b) Supposing that the arithmetic coding algorithm was implemented using a 5-bit register (ie the integer arithmetic it uses is based on the values 0,1,...,31) find what the first few bits of output would be if you started to code the string "33333...". Explain the algorithm, your working and any choices you have to make along the way. [70%]

(c) Arithmetic coding using integer arithmetic should not generate exactly the bit-string you have as your answer to part (a). Explain why not. Discuss whether the bit-string that will be generated is liable to be periodic and whether changing the size of the integer register used makes any substantial changes to behaviour. [20%]

END OF PAPER