COMPUTER SCIENCE TRIPOS Part II (General)
DIPLOMA IN COMPUTER SCIENCE

Wednesday 8 June 2005 1.30 to 4.30

Paper 12 (Paper 3 of Diploma in Computer Science)

Answer **five** *questions.*

Submit the answers in five **separate** *bundles, each with its own cover sheet. On each cover sheet, write the numbers of* **all** *attempted questions, and circle the number of the question attached.*

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

STATIONERY REQUIREMENTS
Script Paper
Blue Coversheets
Tags

1 Data Structures and Algorithms

(*a*) Explain how a Boolean matrix can be used to represent the edges of a finite directed graph whose vertices are numbered 1 to n. [2 marks]

(*b*) Describe Warshall's algorithm to convert the matrix representing a graph to one that represents its transitive closure, and carefully explain why the algorithm works. [6 marks]

(*c*) Outline Floyd's algorithm, without proof of correctness, to find the cost of the cheapest path between any two vertices of a directed graph where the edges carry non-negative costs. [4 marks]

(*d*) It is required to construct a matrix R that encodes a path with the minimum number of edges from any vertex i to any other vertex j. R_{ij} will be zero if no path exists from vertex i to vertex j; otherwise, R_{ij} will hold the vertex number of the next vertex of a minimal path from i to j. Suggest an algorithm to compute R from a given Boolean matrix M. [8 marks]

2 Computer Design

(*a*) What is the difference between a *control hazard* and a *data hazard*? [4 marks]

(*b*) How are data and control hazards handled for the following two processors with their respective pipelines?

The N-105 processor pipeline:

instruction fetch	register fetch, decode, execute memory access and write back

The ARM9 processor pipeline:

instruction fetch	decode	execute	memory access	write back

[8 marks]

(*c*) If a load instruction causes a cache miss, what impact does it have on the pipeline? [3 marks]

(*d*) What is the structure of a TLB (Translation Lookaside Buffer)? [2 marks]

(*e*) What impact does a TLB miss have on the pipeline? [3 marks]

3 Digital Communication I

It is proposed to send information across a fixed delay channel using a simple (window of 1) ARQ protocol with a transmitter timeout of T. That is, if the transmitter does not receive an acknowledgement for a packet within time T of sending the packet, it retransmits.

The delay of the underlying channel is τ, the data rate is B and the packet size is p bits. Bit errors in the channel are independent and packets of size p have a packet error rate of e. Errors in the small acknowledgement packets are rare enough to be discounted in this analysis.

(*a*) What is the expected throughput of the ARQ protocol if e is zero? [4 marks]

(*b*) What is the expected throughput if e is non-zero, but small enough that e^2 is negligibly small? [4 marks]

(*c*) How could a forward error code help the throughput of the ARQ scheme? [2 marks]

(*d*) What is meant by the term *code rate* of a forward error code? [2 marks]

(*e*) What code rate must a code which squared the error rate have in order to improve throughput of the ARQ scheme? [4 marks]

(*f*) If the forward error coder adds delay, how will this affect performance? [4 marks]

4 Distributed Systems

An appropriate structure for large-scale distributed systems is as multiple, independently administered, firewall-protected, domains. Examples are a national health service, a national police service and a global company with worldwide branches. Communication must be supported within and between domains and external services may be accessed. For example, health service domains may all access a national Electronic Health Record service; police service domains may all access a national Vehicle Licensing service.

(*a*) (*i*) Define publish/subscribe communication. [3 marks]

(*ii*) What are the advantages and disadvantages of offering publish/subscribe as the only communication service? [7 marks]

(*b*) (*i*) Define rôle-based access control. [3 marks]

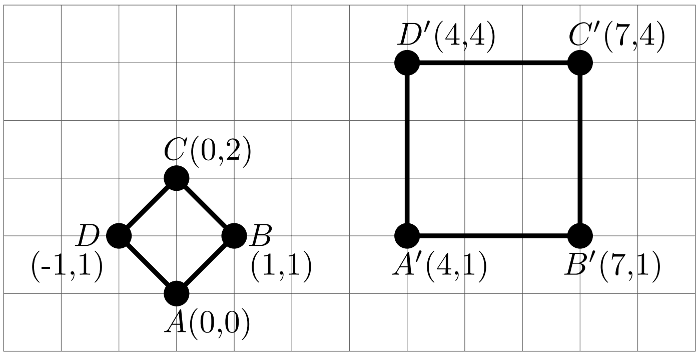
(*ii*) What are the advantages and disadvantages of using rôle names for access control and communication? [7 marks]

Illustrate your discussions by means of examples.

5 Computer Graphics and Image Processing

(*a*) Describe, in detail, an algorithm to clip a straight line against an axis-aligned rectangle. [10 marks]

(*b*) Explain why homogeneous coordinates are used for handling geometric transformations. [3 marks]



(*c*) Give a matrix, or a product of matrices, which will transform the square $ABCD$ to the square $A'B'C'D'$. [4 marks]

(*d*) Show what happens if the same transformation is applied to the square $A'B'C'D'$. [3 marks]

6 Compiler Construction

(*a*) Explain how a parse tree representing an expression can (*i*) be converted into stack-oriented intermediate code and then (*ii*) be translated into simple machine code for a register-oriented architecture (e.g. ARM or IA32) on an instruction-by-instruction basis. Also indicate how this code might be improved to remove push–pop pairs introduced by (*ii*). Your answer need only consider expression forms encountered in the expression:

```
h(a, g(b), c) * 3 + d
```

[12 marks]

(*b*) In Java, expressions are evaluated strictly left-to-right. Consider compiling the function `f` in the following Java class definition:

```
class A
{
    static int a,b;
    void f() { ... <<C>> ... }
    int g(int x) { ... a++; ... }
};
```

Indicate what *both* the intermediate code *and* (improved as above) target code might be for `<<C>>` for the cases where `<<C>>` is:

(*i*) `b = g(7) + a;`

(*ii*) `b = a + g(7);`

(*iii*) `b = (-g(7)) + a;`

(*iv*) `b = a - g(7);`

Comment on any inherent differences in efficiency at both the intermediate code and target code levels.

[8 marks]

7 Comparative Programming Languages

Most large programs that have been written with considerable care and thoroughly checked still seem to contain bugs at a rate of over one per 3000 lines of source code. Systems involving hundreds of millions of lines of code can thus be expected to contain tens of thousands of potentially catastrophic errors.

(*a*) List several kinds of programming errors that can appear in programs and discuss their relative importance in relation to the long-term reliability of a large application program. [7 marks]

(*b*) Suggest potential ways by which programmers may reduce the number of programming errors they make, paying particular attention to language features that might help, extra features in program development systems and possible changes in overall system architecture. [8 marks]

(*c*) In what ways would you expect languages 25 years from now to differ from those that are currently popular? [5 marks]

8 Databases

(*a*) Define the core operators of the relational algebra. [5 marks]

(*b*) Describe *two* differences and *two* similarities between the relational algebra and SQL. [4 marks]

(*c*) Suppose that $S(a,\ b,\ \ldots)$ and $R(a,\ \ldots)$ are relations (the notation indicates that attribute a is in the schema of both S and R, while attribute b is only in the schema of S). Suppose that v is a value; is the following equation always valid?

$$\sigma_{(a=v \text{ or } b=v)}(R \bowtie S) = (\sigma_{a=v}(R)) \bowtie (\sigma_{b=v}(S))$$

If yes, provide a short proof. If no, provide a counter-example. [2 marks]

(*d*) Various *normal forms* are important in relational schema design.

(*i*) Define Third Normal Form (3NF). [3 marks]

(*ii*) Define Boyce-Codd Normal Form (BCNF). [3 marks]

(*iii*) For databases with many concurrent update transactions, explain why schemas in normal form are important for good performance. [3 marks]

9 Numerical Analysis II

The best L_∞ approximation to $f(x) \in C[-1, 1]$ by a polynomial $p_{n-1}(x)$ of degree $n - 1$ has the property that

$$\max_{x \in [-1,1]} |e(x)|$$

is attained at $n + 1$ distinct points $-1 \leq \xi_0 < \xi_1 < \ldots < \xi_n \leq 1$ such that $e(\xi_j) = -e(\xi_{j-1})$ for $j = 1, 2, \ldots n$ where $e(x) = f(x) - p_{n-1}(x)$.

(*a*) Let $f(x) = x^2$. Show, by means of a clearly labelled sketch graph, that the best polynomial approximation of degree 1 is a constant. [3 marks]

(*b*) Now suppose $f(x) = (x + 1)/(x + \frac{5}{3})$ is the function to be approximated over $[-1, 1]$. By sketching the graph, deduce properties of the best linear approximation $p_1(x)$. By differentiating $e(x)$, find $p_1(x)$. [9 marks]

(*c*) Now consider $f(x) = x/(9x^2 + 16)$. Explain why the best approximation over $[-1, 1]$ of degree 2 or less is of the form $p_2(x) = ax$, and sketch the graph to show the extreme values of $e(x)$. Verify that $x = 4/9$ is one of the extreme values and find a. [8 marks]

10 Introduction to Functional Programming

(*a*) Define a polymorphic datatype `'a seq` for lazy sequences, and define functions `head` and `tail` to return the first element and the rest of the sequence respectively. [1 mark each]

(*b*) Define a function `pick` with the following type:

```
'a pick :  'a list -> ('a * 'a list) seq
```

which returns a sequence of pairs `(x, xs)` as in these examples:

(*i*) `pick [1,2,3,4]` returns a sequence with elements `(1,[2,3,4])`, `(2,[1,3,4])`, `(3,[1,2,4])`, `(4,[1,2,3])`.

(*ii*) `pick [1,2,1,2]` returns a sequence with elements `(1,[2,1,2])`, `(2,[1,1,2])`, `(1,[1,2,2])`, `(2,[1,2,1])`.

[4 marks]

(*c*) Define a function `explodeseq` with type

```
explodeseq :  'a list seq -> 'a seq
```

which creates an element in the output sequence from each element in each list of the input sequence. [6 marks]

(*d*) Define a function `implodeseq` with type

```
implodeseq :  int -> 'a seq -> 'a list seq
```

which transforms a sequence of elements into a sequence of lists whose length is specified in the `int` argument. The last list in the output sequence may contain fewer elements. [7 marks]

11 Natural Language Processing

In (1) and (2) below, the words in the sentences have been assigned tags from the CLAWS 5 tagset by a stochastic part-of-speech (POS) tagger:

(1) Turkey_NP0 will_VM0 keep_VVI for_PRP several_DT0 days_NN2 in_PRP a_AT0 fridge_NN1

(2) We_PNP have_VHB hope_VVB that_CJT the_AT0 next_ORD year_NN1 will_VM0 be_VBI peaceful_AJ0

In sentence (1), *Turkey* is tagged as a proper noun (NP0), but should have been tagged as a singular noun (NN1). In sentence (2), *hope* is tagged as the base form of a verb (VVB: i.e., the present tense form other than for third person singular), but should be NN1. All other tags are correct.

(*a*) Describe how the probabilities of the tags are estimated in a basic stochastic POS tagger. [7 marks]

(*b*) Explain how the probability estimates from the training data could have resulted in the tagging errors seen in (1) and (2). [6 marks]

(*c*) In what ways can better probability estimates be obtained to improve the accuracy of the basic POS tagger you described in part (*a*)? For each improvement you mention, explain whether you might expect it to improve performance on examples (1) and (2). [7 marks]

12 Complexity Theory

(*a*) Explain what it means to say that a problem is

(*i*) NP [2 marks]

(*ii*) NP-Complete [2 marks]

(*b*) Define the standard problem 3-SAT and describe how you would take an instance of it and derive an integer n that you would use in any formulae relating to the cost of solving that instance. [3 marks]

(*c*) What is a non-deterministic Turing Machine? Supposing that some computation of such a machine takes N steps, what information needs to be reported to describe exactly how the computation proceeded? In what way is this relevant to the problem of solving an *arbitrary* NP problem? [7 marks]

(*d*) Sketch a proof of Cook's result, that the problem 3-SAT is NP complete. Justify that any transformations you introduce are polynomial. [6 marks]

END OF PAPER