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

Wednesday 4 June 2008 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 cover sheets
Tags

SPECIAL REQUIREMENTS
None

1 Bioinformatics

(a) Parameters of the positional independence of a transcription factor binding site were estimated by the experimental positional nucleotide frequencies shown in the following table:

$$\begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ T & 0.16 & 0.05 & 0.01 & 0.03 & 0.12 & 0.14 \\ C & 0.08 & 0.04 & 0.01 & 0.03 & 0.05 & 0.11 \\ A & 0.68 & 0.11 & 0.02 & 0.90 & 0.16 & 0.51 \\ G & 0.08 & 0.80 & 0.96 & 0.04 & 0.67 & 0.24 \end{pmatrix}$$

Explain what a *logo* is and determine the parameters of the logo graph. Compute the information content of one column. [8 marks]

(b) Discuss the complexity of the algorithm for finding a global alignment between two DNA sequences that have a high degree of similarity. Present an example and analyse it using the following scoring parameters: $+1$ for match, -1 for mismatch, and $d = -1$ for a linear gap penalty. [7 marks]

(c) In modelling a metabolic process, describe the advantages and disadvantages of using a stochastic approach (for example agents) as opposed to using a set of deterministic differential equations. [5 marks]

2 Computer Design

(*a*) The classic MIPS 5-stage pipeline is depicted below.

instruction fetch	decode and register fetch	execute	memory access	write back

(*i*) With reference to the 5-stage pipeline, what are *data hazards* and how can they be resolved to ensure that the programmer's model of sequential execution is always preserved whilst minimising performance impact? [6 marks]

(*ii*) With reference to the 5-stage pipeline, what are *control hazards* and how can they be resolved? [4 marks]

(*b*) If we wanted the above pipeline to mimic two processors running at half the speed, then we could have two copies (A and B) of the register state and keep the existing pipeline. The instruction–fetch stage would alternate between fetching from threads A and B on alternate clock cycles. As a consequence, if instruction fetch was from thread B, then an instruction from thread A would be in decode, B in execute, A in memory access and B in write back.

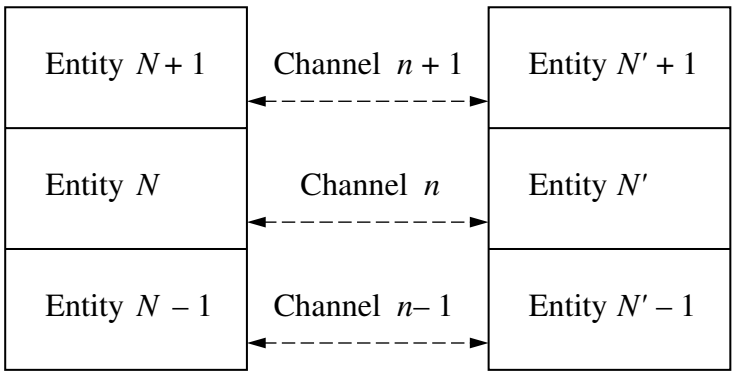
(*i*) For this dual-threaded processor, if branches are performed in the decode stage, does the pipeline exhibit a branch delay slot? [5 marks]

(*ii*) How can the resolution of data hazards be simplified for this dual-threaded processor? [5 marks]

3 Digital Communication I

(a) Describe *five* physical properties of a communications channel. [5 marks]

(b) Consider the figure below. Entities N and N' use an ARQ system.



(i) Explain how the latency of channel $n-1$ can have a direct effect on the capacity of channel n. [6 marks]

(ii) Define *windowing* as it relates to an ARQ system and describe how the capacity of the ARQ system may be improved through its use. [4 marks]

(iii) If an ARQ system is used for an interactive session, the ARQ system can lead to many small packets, each under-full and perhaps sent with significant overhead. Design and describe an algorithm that overcomes the limitation of sending many mostly-empty packets for an interactive session. [5 marks]

4 Distributed Systems

(*a*) It is proposed that persistent, strongly consistent data replicas should be maintained by a widely distributed, open, unstructured process group.

(*i*) Discuss the potential advantages of replication, bearing in mind that strong consistency is required. [2 marks]

(*ii*) Describe algorithms for maintaining strong consistency while retaining at least some of the advantages of replication. Show how your algorithms are robust with respect to concurrency and failure. [8 marks]

(*b*) A distributed conference application provides a shared whiteboard. Each member of the conference has a replica of the whiteboard that is managed by a member of a closed process group. Discuss one approach by which the processes can achieve mutually exclusive access to the whiteboard, prior to propagation of the update to the whole group. [8 marks]

(*c*) Contrast the styles of replica management required for (*a*) and (*b*) above. [2 marks]

5 Computer Graphics and Image Processing

(*a*) Describe in detail an algorithm that returns the minimum distance from a point to a line segment in two dimensions. Ensure that you include all of your assumptions and all necessary mathematical calculations. [7 marks]

(*b*) A quadratic Bézier curve is defined by three points, P_1, P_2, P_3, and a parameter, t:

$$P(t) = (1-t)^2 P_1 + 2t(1-t)P_2 + t^2 P_3,\ 0 \leq t \leq 1$$

Describe an algorithm that draws the quadratic Bézier curve, using straight lines only, to within a tolerance τ. You may use the algorithm from part (*a*) and you may assume that you already have an algorithm for drawing a straight line. [8 marks]

(*c*) Consider the control of detail in a curve that is represented by a sequence of many straight line segments. Describe how Douglas and Pücker's algorithm can be used to remove superfluous points. You may use the algorithm from part (*a*). [5 marks]

6 Compiler Construction

Consider the following grammar for expressions (where `Id` is a terminal symbol representing an identifier resulting from lexical analysis):

```
Expr  ::=  1 | 2 | Id | Expr + Expr | Expr / Expr |
                       Expr ^ Expr | (Expr)
```

(*a*) Explain in what principal respect this grammar is unsatisfactory. [1 mark]

(*b*) Assuming further that `+` is to be left-associative, `^` is to be right-associative and `/` is to be *non-associative* (i.e. `2/2/2` is forbidden but `(2/2)/2` and `2/(2/2)` are allowed), re-write the grammar to reflect this. [4 marks]

(*c*) List the terminal symbols and non-terminal symbols, and count the production rules both in the original grammar and in the grammar in your answer to part (*b*). Indicate the *start symbol* in both grammars. [2 marks]

(*d*) Define a type or types (in C, Java, or ML) suitable for holding an abstract syntax tree resulting from your answer to part (*b*). [2 marks]

(*e*) Give a brief and elementary explanation of the principles of how the grammar resulting from part (*b*) might be used to create a syntax analyser taking a token stream as input (via calls to function `lex()`) and giving as output an abstract syntax tree corresponding to part (*d*). Mention both hand-written and automatically-generated syntax analysers. [8 marks]

(*f*) Summarise any issues related to left- or right-associative operators in the two techniques (in implementing the parser and in constructing the tool) you outlined in part (*e*). [3 marks]

7 Concepts in Programming Languages

(*a*) Write a procedure and a call to it in block-structured pseudocode such that the execution of the procedure under *pass-by-reference* and under *pass-by-value/result* yields different outcomes. Justify your answer.

[7 marks]

(*b*) Explain the meaning of *static* (i.e. compile-time) and *dynamic* (i.e. run-time) type checking.

Compare the advantages and disadvantages of these two approaches to type checking from the point of view of the language designer, the language implementer, and the programmer.

[6 marks]

(*c*) Explain how *objects* can be simulated in SML, giving an example.

Does it follow that SML, together with its module system, is an object-oriented programming language? Why?

[7 marks]

8 Databases

(*a*) What is the difference between a *key* and a *functional dependency*? [3 marks]

(*b*) The schema $R(A,\ B,\ C,\ D,\ E)$ has the following functional dependencies.

$$\begin{aligned} A,\ B &\rightarrow C \\ B,\ C &\rightarrow D \\ C,\ D &\rightarrow E \\ D,\ E &\rightarrow A \end{aligned}$$

(*i*) What are all of the keys of R? [3 marks]

(*ii*) Which functional dependencies violate Boyce–Codd Normal form (BCNF)? [3 marks]

(*iii*) Which functional dependencies violate Third Normal form (3NF)? [3 marks]

(*iv*) Find a lossless-join decomposition of R into BCNF relations. [8 marks]

9 Mathematics for Computation Theory

(*a*) Let $f : A \to B$ be a function with domain A and range B. Show that the relation R on A defined by

$$(x, y) \in R \quad \Leftrightarrow \quad f(x) = f(y)$$

is an equivalence relation. [4 marks]

(*b*) A *partition* of a set A is a set $\mathcal{A}$ of disjoint subsets of A such that $A = \bigcup \mathcal{A}$, *and*

$$B, C \in \mathcal{A} \quad \Rightarrow \quad (B = C) \quad \vee \quad (B \cap C) = \emptyset$$

Let $g(n, r)$ be the number of partitions of a set A having n elements into r subsets, where $1 \leqslant r \leqslant n$. If $1 < r < n$, show that

$$g(n, r) \quad = \quad r\ g(n-1, r) + g(n-1, r-1). \qquad \text{[7 marks]}$$

(*c*) Using the above formula, or otherwise, evaluate $g(n, r)$ in the cases:

(*i*) $r = 2$ [4 marks]

(*ii*) $r = (n - 1)$ [5 marks]

10 Computation Theory

(*a*) *The Halting Problem for register machines is unsolvable.* State, without proof, a precise form of this result. [3 marks]

(*b*) Let the computation by program c on data d be represented by the natural number k that codes the pair (c, d). By considering the set $H(k)$ of the HALTing computations represented by codes $k' < k$, show that there is an increasing total function $h(k)$ which *grows too fast* to be computable. [6 marks]

(*c*) Given $h : \mathbb{N} \to \mathbb{N}$ with the above property

$$\begin{aligned} \text{let} \quad f(k) &= h(k) + k \\ \text{and} \quad g(x) &= \sup\{k : f(k) \leqslant x\}. \end{aligned}$$

Then $f : \mathbb{N} \to \mathbb{N}$ is strictly increasing, and $g : \mathbb{N} \to \mathbb{N}$ satisfies

$$g(f(k)) = k, \quad g(x) < k \quad \text{for all } x < f(k).$$

Show that g *grows too slowly* to be computable in the following sense:

given $G : \mathbb{N} \to \mathbb{N}$ such that

(*i*) $\{G(n) : n \in \mathbb{N}\}$ is unbounded

(*ii*) $G(n) \leqslant g(n)$ for all $n \in \mathbb{N}$

then $G(n)$ is *not* computable. [11 marks]

11 Software Engineering and Design

The term "business-critical" is sometimes used, by analogy to "safety-critical", to describe a system in which software failure could lead to failure of the whole business. If this analogy were to be taken seriously, many of the design precautions used for safety-critical software could also be used in the design of business-critical software.

Describe *five* different aspects of the analysis and design process that are specific to safety-critical software, and for each of these, explain how they might be adapted to business-critical software. [4 marks each]

12 Complexity Theory

(a) Give a precise definition of what it means for one decision problem to be polynomial-time reducible to another. [3 marks]

(b) Consider the following two decision problems:

HamCycle: Given a graph $G = (V, E)$ does it contain a cycle that visits every vertex exactly once?

HamPath: Given a graph $G = (V, E)$ and two distinguished vertices $s, t \in V$, is there a simple path in G that starts at s, ends at t and visits every other vertex exactly once?

Show that **HamCycle** is polynomial-time reducible to **HamPath**. [8 marks]

(c) The following decision problem is known to be solvable in polynomial time:

EulerCycle: Given a graph $G = (V, E)$ does it contain a cycle that visits every edge exactly once?

What can you conclude about the truth of the following statements? Justify your answers.

(i) **EulerCycle** is polynomial-time reducible to **HamCycle**. [3 marks]

(ii) **EulerCycle** is polynomial-time reducible to **HamPath**. [3 marks]

(iii) **HamPath** is polynomial-time reducible to **EulerCycle**. [3 marks]

END OF PAPER