COMPUTER SCIENCE TRIPOS Part II

Wednesday 4 June 2008 1.30 to 4.30

PAPER 8

*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:

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

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 VLSI Design

(*a*) Sketch a transistor-level circuit for the function $\overline{A \cdot B + C \cdot D}$ in static CMOS. [4 marks]

(*b*) Annotate the circuit to indicate the widths of the transistors required to give rise and fall times equal to a minimal, balanced inverter. Assume that p-channel transistors have γ times the resistance of n-channel transistors when conducting. [4 marks]

(*c*) Calculate the logical effort and parasitic delay for the circuit. [4 marks]

(*d*) Sketch a stick diagram for the circuit, arranged for reasonably compact layout. Assume two layers of metal with power and ground routed in parallel tracks on the second layer, and arrange for inputs and the output to be available outside the power rails. [4 marks]

(*e*) Estimate the size of your layout, assuming a separation of 8λ between the centre-lines of parallel metal tracks. [4 marks]

3 Digital Communication II

(*a*) What are the pros and cons of *distance vector* versus *link state* routing protocols? Give examples derived from protocols in use today. [10 marks]

(*b*) Where are hybrid schemes employed and why? [5 marks]

(*c*) Again using examples, discuss the issues in extending routing to support multicast. [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 Comparative Architectures

(*a*) Why is it important to concentrate on improving the common case (e.g. the most commonly used operations and resources) when designing a microprocessor? [4 marks]

(*b*) What is the major difference between a very long instruction word (VLIW) processor and a dynamically-scheduled superscalar processor? What impact does this have on the complexity of the implementation in each case? [4 marks]

(*c*) When designing a VLIW processor, why might variable-length instruction bundles be preferred over fixed-length instructions? [4 marks]

(*d*) Some VLIW processors contain additional hardware to permit memory reference speculation.

(*i*) What optimisation does memory reference speculation permit? [4 marks]

(*ii*) Briefly describe the additional hardware required to support this type of speculation. [4 marks]

6 Computer Vision

(*a*) Consider the *eigenface* algorithm for face recognition in computer vision.

(*i*) What is the rôle of the database population of example faces upon which this algorithm depends? [3 marks]

(*ii*) What are the features that the algorithm extracts, and how does it compute them? How is any given face represented in terms of the existing population of faces? [4 marks]

(*iii*) What are the strengths and the weaknesses of this type of representation for human faces? What invariances, if any, does this algorithm capture over the factors of perspective angle (or pose), illumination geometry, and facial expression? [4 marks]

(*iv*) Describe the relative computational complexity of this algorithm, its ability to learn over time, and its typical performance in face recognition trials. [3 marks]

(*b*) In a visual inference problem, we have some data set of observed features x, and we have a set of object classes $\{C_k\}$ about which we have some prior knowledge. Bayesian pattern classification asserts that:

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}$$

Explain the meaning of, and give the name for, each of these three terms:

$P(C_k|x)$
$P(x|C_k)$
$P(C_k)$ [3 marks]

(*c*) Define the concept of *reflectance map* $\phi(i, e, g)$ and define the three variables i, e, and g on which it depends. [3 marks]

7 Advanced Graphics

(*a*) Place four control points P_1, P_2, P_3, P_4 in a square. For each of the following knot vectors, for the quadratic B-spline ($k = 3$), sketch (*i*) the four basis functions and (*ii*) the B-spline curve defined by the four control points and four basis functions, marking the location of the knots and the value of t at each knot.

(α) $[1, 2, 3, 4, 5, 6, 7]$

(β) $[1, 2, 3, 3, 4, 5, 6]$

(γ) $[1, 2, 3, 3, 3, 4, 5]$

[12 marks]

(*b*) Describe, in detail, an algorithm to find the intersection point between an arbitrary ray and an arbitrary triangle in 3D. Ensure that you define all parameters. [8 marks]

8 Optimising Compilers

(*a*) Sometimes evaluating expressions may be partially or wholly redundant in that they have been previously evaluated on some or all of the program paths leading to them.

(*i*) Outline the theory of available expressions, including dataflow equations and how to compute their solution. Also give a brief explanation of how to use this solution to remove common-subexpressions. How does the idea of either form of redundant computation relate to the notion of common-subexpression? [7 marks]

(*ii*) Give an example of a redundant computation that is not removed by the technique you give in part (*i*). [2 marks]

(*b*) Consider an intra-procedural dataflow analysis for security. Variables may hold high-security (e.g. a PIN) or low-security (e.g. a counter) values. Program constants are low-security, and on function entry only variables in the set H are high-security. Security flows through direct dataflow: the result of an assignment is assumed to be high-security if a variable on the right-hand side may hold a high-security value.

(*i*) Design a dataflow analysis that calculates, for each node n in a flowgraph, the set of variables that *may* hold a high-security value at n. Have you defined a forward analysis or backward analysis? How is your dataflow analysis implemented, noting particularly initialisation of any iteration? [7 marks]

(*ii*) Give an informal argument as to why your dataflow analysis is safe or an example of why it is not—in either case discussing reasons or interesting cases. For this purpose treat an analysis as being safe if it is impossible to write a function body that (1) implements the identity function and (2) has the property that the output variable is analysed as low-security on exit even though the input variable is high-security (a member of H) on entry. [4 marks]

9 Information Theory and Coding

(*a*) (*i*) A variable-length, uniquely decodable code that has the prefix property and whose N binary code word lengths are $n_1 \leq n_2 \leq n_3 \leq \cdots \leq n_N$ must satisfy what condition with these code word lengths? (Give an expression for the condition, and its name, but do not attempt to prove it.) [3 marks]

(*ii*) Construct an efficient, uniquely decodable binary code, having the prefix property and having the shortest possible average code length per symbol, for an alphabet whose five letters appear with these probabilities:

Letter	A	B	C	D	E
Probability	1/2	1/4	1/8	1/16	1/16

[3 marks]

(*iii*) How do you know that, on average, for samples drawn from this alphabet, your code uses the shortest possible code length per symbol? Demonstrate numerically that your code satisfies this optimality condition. [3 marks]

(*b*) (*i*) Explain how autocorrelation can remove noise from a signal that is buried in noise, recovering a clean signal. For what kinds of signals, and for what kinds of noise, will this work best, and why? What class of signals can be recovered perfectly by autocorrelation? Begin your answer by writing down the integral that defines the autocorrelation of a signal $f(x)$. [3 marks]

(*ii*) Some sources of noise are additive (the noise is just superimposed onto the signal), but other sources of noise are multiplicative in their effect on the signal. For which type would the autocorrelation clean-up strategy be more effective, and why? In the case of additive noise where noise and signal occupy different frequency bands, what other strategy could allow recovery of a clean signal? [3 marks]

(*c*) (*i*) If a continuous signal $f(t)$ is *modulated* by multiplying it with a complex exponential wave $\exp(i\omega t)$ whose frequency is ω, what happens to the Fourier spectrum of the signal? Name a very important practical application of this principle, and explain why modulation is a useful operation. How can *demodulation* then recover the original signal? [3 marks]

(*ii*) Which part of the 2D Fourier Transform of an image, the amplitude spectrum or the phase spectrum, is indispensable in order for the image to be intelligible? Describe a demonstration that proves this. [2 marks]

10 Security

The One Laptop Per Child project aims to supply millions of rugged low-cost laptops to children in less-developed countries. The machines run Linux, have 2 Gb Flash rather than a hard disk, and have a wireless LAN capability that may be used either in the conventional way or to set up *ad-hoc* peer-to-peer networks.

Your task is to design the security policy for these laptops. If the project is to supply standard security software with each machine, what should it try to do, and how? [20 marks]

11 Digital Signal Processing

(*a*) What is the Fourier transform of a rectangular pulse of amplitude A and duration $d > 0$, centred around $t = 0$? [4 marks]

(*b*) Calculate the Fourier transform of the triangular pulse

$$\Lambda(t) = \begin{cases} 1 - |t|, & \text{for } |t| < 1 \\ 0, & \text{otherwise} \end{cases}$$

[Hint: Think of $\Lambda(t)$ as the result of a convolution.] [4 marks]

(*c*) A 2 kHz sine wave is sampled at 12 kHz. The resulting values are later converted back into a continuous signal using *linear interpolation.*

(*i*) At what other frequencies besides 2 kHz is there signal energy in the resulting continuous waveform? [4 marks]

(*ii*) Consider among those other components the one with the lowest frequency. By what factor is its voltage lower compared with the 2 kHz component? [4 marks]

(*iii*) Your colleague records with a PC soundcard at 44.1 kHz sampling frequency 1024 samples of the continuous waveform, loads these into MATLAB as vector `x` and then attempts to plot an amplitude spectrum with the command

```
plot(real(fft(x)));
```

Name *two* problems that need to be fixed in this command before the resulting plot is likely to agree with the result of (*ii*). [4 marks]

12 Computer Systems Modelling

(*a*) Explain what is meant by a birth–death model with birth rates λ_i and death rates μ_i in states $i = 0, 1, \ldots$. You should include in your explanation the necessary probabilistic assumptions. [4 marks]

(*b*) Write down the *detailed balance* equations for an equilibrium distribution, p_i, of being in state i in the birth–death model. Use these equations to determine the p_i and clarify when such an equilibrium distribution exists. [4 marks]

(*c*) Consider the $M/M/m/m$ model of a loss system with m servers. Describe how this system can be used to model the behaviour of a telephone link consisting of C circuits with an arrival rate of λ calls per second and a mean holding time of $\frac{1}{\mu}$ seconds. [4 marks]

(*d*) Use your general results from part (*b*) to derive the equilibrium distribution of the number of free circuits on a telephone link and hence deduce *Erlang's formula* for the probability that there are no free circuits available. What are the conditions for the equilibrium distribution to exist? [4 marks]

(*e*) Comment on any numerical problems that could arise in calculating Erlang's formula when C is large. How might you overcome these difficulties? [4 marks]

13 Types

(*a*) Explain what is meant by the relation of *specialisation*, $\sigma \succ \tau$, between Mini-ML type schemes σ and Mini-ML types τ. How is $\succ$ used in the Mini-ML type system? [4 marks]

Assuming α_1 and α_2 are distinct type variables, which of the following are valid instances of specialisation?

(*i*) $\forall \alpha_1, \alpha_2(\alpha_1 \to \alpha_2) \succ (\alpha_1 \to \alpha_1) \to \alpha_1$

(*ii*) $\forall \alpha_1(\alpha_1 \to \alpha_2) \succ (\alpha_1 \to \alpha_1) \to \alpha_1$

(*iii*) $\forall \alpha_1(\alpha_1 \to \alpha_2) \succ (\alpha_2 \to \alpha_2) \to \alpha_2$

(*iv*) $\forall \alpha_1(\alpha_1 \to \alpha_1) \succ (\alpha_1 \to \alpha_1) \to \alpha_2$

[6 marks]

(*b*) Extending Mini-ML with fixed-point expressions $\mathtt{fix}\, x(M)$, consider the following typing rules:

$$\text{(mono-fix)} \quad \frac{\Gamma, x : \forall\{\}(\tau) \vdash M : \tau}{\Gamma \vdash \mathtt{fix}\, x(M) : \tau} \quad \text{if } x \notin \mathit{dom}(\Gamma)$$

$$\text{(poly-fix)} \quad \frac{\Gamma, x : \forall A(\tau) \vdash M : \tau}{\Gamma \vdash \mathtt{fix}\, x(M) : \tau} \quad \text{if } x \notin \mathit{dom}(\Gamma) \text{ and } A = \mathit{ftv}(\tau) - \mathit{ftv}(\Gamma)$$

(where as usual $\mathit{ftv}(-)$ indicates the set of free type variables in $-$). Write $\Gamma \vdash_{\text{mono}} M : \tau$ (respectively $\Gamma \vdash_{\text{poly}} M : \tau$) if $\Gamma \vdash M : \tau$ is provable in the Mini-ML type system extended with the rule (mono-fix) (respectively with the rule (poly-fix)). Let $M = \mathtt{fix}\, x(\lambda y((x\, x) y))$. State, with justification, which of the following hold for some type τ.

(*i*) $\{\} \vdash_{\text{mono}} M : \tau$

(*ii*) $\{\} \vdash_{\text{poly}} M : \tau$

[10 marks]

14 Denotational Semantics

(*a*) Describe the properties a function between two cpos must have to be continuous. [2 marks]

(*b*) Let D_1, D_2 and E be cpos. Prove that a function $h : D_1 \times D_2 \to E$ is continuous if it is continuous in each argument separately. [You may assume standard properties of least upper bounds provided you state them clearly.] [4 marks]

(*c*) Let $\mathbb{O}$ be the cpo with two elements $\bot \sqsubseteq \top$. For a cpo E and $e \in E$, define the function $g_e : E \to \mathbb{O}$ by

$$g_e(x) = \begin{cases} \bot & \text{if } x \sqsubseteq e \\ \top & \text{if } x \not\sqsubseteq e \end{cases}$$

Show g_e is continuous. [4 marks]

(*d*) As an example of the definition in part (*c*) above, let $E = \mathbb{B}_\bot \times \mathbb{B}_\bot$, where $\mathbb{B} = \{\mathit{true}, \mathit{false}\}$, and consider $g_{(\mathit{false},\mathit{false})} : E \to \mathbb{O}$. Show that

$$g_{(\mathit{false},\mathit{false})}(x, y) = \top \text{ iff } x = \mathit{true} \text{ or } y = \mathit{true}$$

[2 marks]

(*e*) Let $f : D \to E$ be a function between cpos D and E. Show

$$f \text{ is continuous iff } \forall e \in E. \quad g_e \circ f \text{ is continuous}$$

[You may assume that the composition of continuous functions is continuous. It is suggested that for the "if" direction of the proof, you argue by contradiction.] [8 marks]

15 Topics in Concurrency

(*a*) A *simulation* between CCS terms is defined to be a binary relation S between CCS terms such that whenever $(t, u) \in S$ for all actions a and terms t'

$$t \xrightarrow{a} t' \Rightarrow \exists u'.\ u \xrightarrow{a} u' \ \&\ (t', u') \in S$$

Write $t \leq u$ iff there is a simulation S for which $(t, u) \in S$. Consider the following fragment of Hennessy–Milner logic:

$$A ::= \langle a \rangle A \mid \bigwedge_{i \in I} A_i$$

where a is an action of CCS and I is a set. In fact,

> $t \leq u$ iff for all assertions A in the fragment whenever t satisfies A then so does u.

(*i*) Explain briefly the strategy you would use to prove the "only if" direction of the fact above; state clearly any induction hypothesis you would use. [3 marks]

(*ii*) Prove the "if" direction. [7 marks]

(*b*) Describe a Petri net semantics for the following fragment of CCS:

$$t ::= rec\, x\, s \mid t_1 \parallel t_2 \mid t \setminus b$$

in which

$$s ::= \alpha.x \mid \alpha.s \mid s_1 + s_2$$

where α ranges over the actions of CCS, b over non-τ actions and x over process variables.

A diagrammatic account suffices, though you should make clear the form of labelled Petri net you are using and its "token game." Although no proof is needed, your semantics should represent the independence of actions in a parallel composition and agree with the usual transition semantics of CCS. [10 marks]

16 Specification and Verification I

(*a*) What is the difference between partial and total correctness? Illustrate your answer using the `WHILE`-Rule. [4 marks]

(*b*) What are derived rules? Give an example of a derived rule, together with its derivation. [4 marks]

(*c*) If $\{P\}\ C\ \{Q\}$ is a theorem of Hoare logic and if C is correctly annotated then will the verification conditions necessarily be provable? Justify your answer. [4 marks]

(*d*) Would a Java program that translates Hoare formulae to semantically equivalent higher-order logic formulae be a deep embedding or a shallow embedding? Justify your answer. [4 marks]

(*e*) Describe the meaning of the dynamic logic formulae $[c]q$ and $<c>q$. For deterministic commands, describe how partial and total Hoare logic correctness specifications can be formulated in terms of dynamic logic. [4 marks]

END OF PAPER