CST2
COMPUTER SCIENCE TRIPOS Part II

Tuesday 10 June 2025 13:30 to 16:30

COMPUTER SCIENCE Paper 9

*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
Approved calculator permitted

1 Advanced Computer Architecture

(*a*) Conventional instruction sets specify an instruction's operands by using register names. We could alternatively specify an instruction's operands by inter-instruction distance, that is, indicating the required operand by counting backwards to the instruction that generated it.

(*i*) Describe one possible advantage of this approach. [3 marks]

(*ii*) If we specified an instruction's operands in this way we could still retain the register file. Describe a simple way to determine the destination register for each instruction that would mean it would not be encoded in the instruction. [3 marks]

(*iii*) In what cases would it be difficult to specify operands using the scheme described? [6 marks]

(*b*) Precisely what invariants does a cache coherence protocol guarantee? [3 marks]

(*c*) Imagine a multi-core system with a directory-based cache-coherence protocol.

(*i*) What are the benefits of a directory-based cache coherence protocol over a snooping protocol? [2 marks]

(*ii*) Why might an L3 cache have more tags and directory entries than cache lines, that is, be a non-inclusive cache that maintains an inclusive directory? [3 marks]

2 Bioinformatics

(a) Genetic toggle switches are systems of two proteins that repress each other's production. Such systems typically exhibit bistability and switching behaviours. Let's suppose that gene a produces an mRNA that translates into a protein A that inhibits gene b, which produces an mRNA that translates into a protein B, that inhibits gene a.

(i) Describe how the Gillespie algorithm models this circuit. [8 marks]

(ii) What factors affect the accuracy or robustness of the model output? [2 marks]

(b) Contrast Progressive Multiple sequence Alignment (PMA) with the Smith-Waterman (S/W) algorithm, using a common example. (The example needs to contain at least three sequences, at least 3-4 bases each.) Is a scoring matrix important in either algorithm? [4 marks]

(c) You are sequencing a new pathogenic bacterium.

(i) How would you proceed to carry out phylogenetic analysis to characterise the taxonomy of the bacterium? [3 marks]

(ii) In practice, many researchers compare results from multiple methods to see which species are robustly supported, rather than relying on a single approach. If you use parsimony and distance methods (for example neighbour joining) would you get the same information or different? [3 marks]

3 Business Studies

Giles Murchiston has returned to the Lab to lead a research group developing a technology to give reliable research advice to researchers using an AI chatbot.

After much development Giles decides to productise the research.

(*a*) Name and describe five types of intellectual property that would be relevant to protecting an AI chatbot product. [5 marks]

(*b*) Giles is considering whether to release his technology as an open-source code library or as a chatbot app.

Describe five productisation decisions Giles would have to consider and how they would differ under each option. [15 marks]

4 Cryptography

(*a*) List *six* properties that an algebraic group should have to be usable for Diffie–Hellman key exchanges. [6 marks]

(*b*) Let $T : A^8 \to A^4$ be a new collision-resistant compression function approved for use in Tripos papers, where $A = \{\texttt{a}, \ldots, \texttt{z}, \texttt{0}, \ldots, \texttt{9}, \texttt{=}, \texttt{\&}\}$ is the "base38" alphabet used.

(*i*) Assuming a Tripos student with pocket calculator can evaluate T once per minute, and assuming all students have a brain with unlimited memory and instantaneous recall time, how many hours will it roughly take until at least half of all students can be expected to each have independently found a collision $T(x) = T(y)$ with $x \neq y$? [2 marks]

(*ii*) Use T to define a collision-resistant hash function $H : A^* \to A^4$, such that the security proof for the Merkle–Damgård construction can be applied. Describe your padding scheme and list the input blocks fed into T when you evaluate $H($"`love&peace`"$)$. [6 marks]

(*iii*) Consider an ATM that receives from a bank computer authorization responses of the form (M, C), such as

$$M = \text{"}\texttt{txn=491\&pincheck=0\&limit=0}\text{"}, \quad C = H(K \| M)$$

where $K \in A^8$ is the private key shared between the bank and the ATM, and H is as in Part (*b*)(*ii*).

After recalculating and checking C, the ATM splits M into fields separated by "`&`", and then executes any variable assignments it encounters in such fields from left to right, ignoring fields that do not form an assignment. The above M confirms that the PIN provided for transaction 491 was incorrect and that the cardholder is therefore authorized to receive up to £0 in cash.

Mallory has intercepted the line between the ATM and the bank computer and can read (M, C) and replace it with a modified message (M', C'). She would like to withdraw cash without knowing the PIN. Show how she can form a message M' that ends in "`&pincheck=1&limit=1000`" and how she can calculate for that M' a matching tag $C' = H(K \| M')$ without knowing K. [6 marks]

5 Denotational Semantics

We consider the language called System T, whose types and terms are respectively given by the following grammars:

$$\tau := \mathtt{nat} \mid \tau \to \tau$$
$$t := \mathtt{0} \mid \mathtt{succ}(t) \mid \mathtt{iter}(t,t,t) \mid x \mid \mathtt{fun}\, x\colon \tau.\ t \mid t\ t$$

Contexts, as in PCF, are partial maps from variables to types. Typing and operational semantics is the same as in PCF, except for iteration, which is as follows:

$$\frac{\Gamma \vdash n : \mathtt{nat} \quad \Gamma \vdash t_0 : T \quad \Gamma \vdash t_{\mathtt{succ}} : T \to T}{\Gamma \vdash \mathtt{iter}(n, t_0, t_{\mathtt{succ}}) : T}$$

$$\frac{n \Downarrow_{\mathtt{nat}} \mathtt{0} \quad t_0 \Downarrow_\tau v}{\mathtt{iter}(n, t_0, t_{\mathtt{succ}}) \Downarrow_\tau v} \qquad \frac{n \Downarrow_{\mathtt{nat}} \mathtt{succ}(n') \qquad t_{\mathtt{succ}}\ \ \mathtt{iter}(n', t_0, t_{\mathtt{succ}}) \Downarrow_\tau v}{\mathtt{iter}(n, t_0, t_{\mathtt{succ}}) \Downarrow_\tau v}$$

We keep the PCF rules for the operational semantics of values, successor and function application.

In System T, we only include bounded iteration, so all functions are total. The goal of this question is to give a denotational semantics to System T using **sets** and **total functions** rather than domains to reflect this.

(a) Give the PCF (and System T) typing rules for 0, `succ`, variables, `fun` and application, and the operational semantic rules for values, successor and application. [3 marks]

(b) Give a denotation $[\![\cdot]\!]$ to System T types and contexts, such that if τ is a type then $[\![\tau]\!]$ is a set. [2 marks]

(c) Given the answer to (b), state what should be the interpretation of a System T term t such that $\Gamma \vdash t : \tau$. [2 marks]

(d) Give a denotational semantics $[\![\cdot]\!]$ for (well-typed) System T terms. Justify that this denotation is well-defined. [5 marks]

(e) State what it means for this denotation to be sound. Show that this is indeed the case. You can freely assume that the denotation is substitutive, provided you clearly state that property. [5 marks]

(f) Let $\overline{0} \in \mathbb{N} \to \mathbb{N}$ be the constant 0 function, and

$$\begin{array}{rlcll} \text{is\_zeroes} \in & (\mathbb{N} \to \mathbb{N}) & \to & \mathbb{N} & \\ & \overline{0} & \mapsto & 0 & \\ & f & \mapsto & 1 & \text{otherwise} \end{array}$$

Do you think is_zeroes is definable with respect to the semantics from part (d)? Explain informally why, or why not, and what could be a strategy to prove this fact. You do *not* need to give a full proof. [3 marks]

6 Hoare Logic and Model Checking

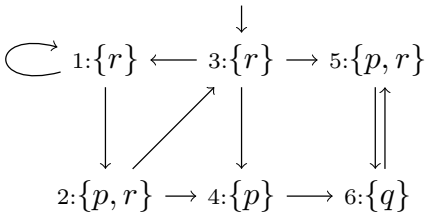
Consider the temporal logic CTL over atomic propositions $p \in \mathit{AP}$:
$\psi \in \mathsf{StateProp} ::= \bot \mid \top \mid \neg\psi \mid \psi_1 \wedge \psi_2 \mid \psi_1 \vee \psi_2 \mid \psi_1 \to \psi_2 \mid p \mid \mathsf{A}\ \phi \mid \mathsf{E}\ \phi$,
$\phi \in \mathsf{PathProp} ::= \mathsf{X}\ \psi \mid \mathsf{F}\ \psi \mid \mathsf{G}\ \psi \mid \psi_1\ \mathsf{U}\ \psi_2$

(a) Specify the following properties as CTL formulae over $\mathit{AP} = \{p, q\}$.

(i) If a state satisfying p can be reached, then there is a path along which q holds until p does. [2 marks]

(ii) All next states have a path that traverses only states from where q cannot be reached. [3 marks]

(b) Consider a temporal model M over atomic propositions $\mathit{AP} = \{p, q, r\}$ with states 1, ..., 6, initial state 3 and transitions and state labelling as shown in the diagram (for example, in state 5, atomic propositions p and r hold).

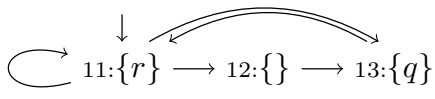


Informally describe the meaning of each of the following CTL formulae over AP and explain whether or not they hold in the model.

(i) $\mathsf{AG}((\mathsf{EX}p) \to (\mathsf{AF}p))$ [3 marks]

(ii) $\mathsf{E}(r\mathsf{U}(\mathsf{AX}(q \wedge \mathsf{AX}r)))$ [3 marks]

(c) Let M be the model from (b) over atomic propositions $\mathit{AP} = \{p, q, r\}$ and M' the temporal model over atomic propositions $\mathit{AP}' = \{q, r\}$ with states 11, 12, and 13, initial state 11, and transitions and labelling as shown below. Prove that M' simulates M, explaining your steps.



[6 marks]

(d) Consider CTL* formula $\psi_1 = \mathsf{A}(\mathsf{G}p \vee \mathsf{F}q)$ and CTL formula $\psi_2 = \mathsf{AG}(p \vee \mathsf{AF}q)$, both over atomic propositions $\mathit{AP} = \{p, q\}$. Formally define a temporal model over AP that shows that ψ_1 and ψ_2 are not equivalent. Explain why your temporal model satisfies one of the formulae but not the other. [3 marks]

7 Information Theory

(*a*) (*i*) Give two advantages and one disadvantage of Arithmetic coding over Huffman coding. [3 marks]

(*ii*) Describe the possible outcomes of a single bit transmission error when decoding a stream of data using a Huffman code. Discuss whether the decoder can detect any errors that occur. How would your answer change for an Arithmetic coding scheme? [8 marks]

(*iii*) Explain why Large Language Models can be used with Arithmetic coding to achieve better text compression. What are the disadvantages of this approach? [2 marks]

(*b*) (*i*) Derive the relationship between k and p for a Hamming code where each block has k data bits and p parity bits. [3 marks]

(*ii*) A B-interleaved coding scheme transmits B consecutive Hamming codewords by interleaving their bits. For example, a 2-interleaved system using a (7,4) Hamming code transmit adjacent codewords 0000000 and 1111111 as 01010101010101. Where might this scheme have an advantage over a conventional transmission of each codeword? Discuss the choice of B. [4 marks]

8 Machine Learning and Bayesian Inference

You have a two-class classification problem, with classes $\{c_+, c_-\}$. Rather than making a simple prediction of which class an example $\mathbf{x}$ falls into, you wish to make two kinds of judgement:

(a) $\mathbf{x}$ is predicted to be in class c_i, where $i \in \{+, -\}$.

(b) $\mathbf{x}$ is *weakly* predicted to be in class c_i, where $i \in \{+, -\}$.

Predicting class membership according to (a) results in a loss of 0 if the prediction is correct and a loss of 1 if in error. Weakly predicting class membership according to (b) results in a loss of θ_1 if correct and a loss of $1 - \theta_2$ if in error. We assume $0 \leq \theta_i \leq 1$ where $i \in \{1, 2\}$. Your aim is to design a *Bayes decision rule* for this problem.

(a) Denoting by $\Pr(C|\mathbf{x})$ the conditional distribution of the class, give a definition of *conditional risk*. [3 marks]

(b) Let $p = \Pr(c_+|\mathbf{x})$. Write down expressions for the *conditional risks* for the actions described. [4 marks]

(c) What constraint should be placed on the values θ_1 and θ_2 such that the options for weak predictions will be relevant in applying the Bayes decision rule? [4 marks]

(d) Derive the Bayes decision rule for this problem, assuming that the constraints derived in Part (c) are met. [4 marks]

You now wish to add a third possibility: 'I *decline* to make a prediction for $\mathbf{x}$'. Declining to make a prediction always has a loss of θ_3, where $0 \leq \theta_3 \leq 1$.

(e) What additional constraint is needed on θ_3 such that it will be relevant in applying the Bayes decision rule? [2 marks]

(f) What modification is necessary to your answer to Part (d) in order to include the option to decline in the Bayes decision rule? [3 marks]

9 Optimising Compilers

Consider the following abstract syntax for a language $\mathcal{L}$ whose types are integers and functions:

$$e ::= x \mid \lambda x.e \mid e_1 e_2 \mid \mathcal{G}(x) \mid \mathcal{G}(x) := e \mid \texttt{if } e_1 \texttt{ then } e_2 \texttt{ else } e_3 \mid \texttt{let } x = e_1 \texttt{ in } e_2$$

where x ranges over variable names, $\mathcal{G}(x)$ reads from global variable x and $\mathcal{G}(x) := e$ evaluates e, writes its result to global variable x and itself evaluates to the value of e.

(a) Provide inference rules for a type-and-effect system for $\mathcal{L}$, where effects are a subset of $\{R_x, W_x \mid x \text{ is a global variable}\}$. [7 marks]

(b) Show how the rules from part (a) assign a type and effect(s) to the following expressions:

(i) $\mathcal{G}(y) := \mathcal{G}(x)$ [1 mark]

(ii) $\texttt{let } f = \lambda x.\mathcal{G}(y) := x \texttt{ in } f\ \mathcal{G}(x)$ [3 marks]

(iii) $\texttt{if } \mathcal{G}(x) \texttt{ then } \lambda x.\mathcal{G}(y) := x \texttt{ else } \lambda x.x$ [3 marks]

(c) Each global variable has its own lock that needs to be taken before reading or writing to it, which is achieved in $\mathcal{L}$ with a new construct:

$$e ::= \texttt{synchronised } e$$

that provides mutual exclusion by taking the locks required for the evaluation of e before e is executed and unlocking them afterwards. The type-and-effect system can be used to help identify which locks should be taken at each `synchronised` expression. Extend your type-and-effect system with an inference rule for this new construct that can help with this analysis and explain this rule. [*Note:* you do not need to provide any inference rules for the locking and unlocking operations themselves.] [4 marks]

(d) Discuss the relative merits of using effect sets compared to effect sequences when generating code to take and release locks for the construct in part (c). [2 marks]

10 Principles of Communications

(*a*) How might fibbing work with IP multicast routing?

Recall that fibbing is a hybrid of distributed and centralised routing that injects advertisements for virtual nodes and links using the intra-domain routing protocol. [10 marks]

(*b*) Describe with examples what adding mobility of end systems does to Traffic Engineering. Assume a simple hill-climbing system for balancing traffic across various paths in the network. [10 marks]

11 Quantum Computing

Quantum error correction.

(a) Define the 3 qubit bit flip code. What is the logical state encoded in this code? [2 marks]

(b) What is a logical X operator for this code? What are the different possible logical Z operators for this code? [3 marks]

(c) Prove or disprove: when a general single qubit state $a|0\rangle + b|1\rangle$ is encoded using the 3 qubit bit flip code, we will violate the no cloning theorem because the encoding circuit of the code will copy the qubit's state onto two others. [2 marks]

(d) The following defines a 4-qubit error detecting code that encodes the state of two data qubits using four physical qubits. The codewords corresponding to the four basis states of the two data qubits are:

$$|00\rangle \rightarrow (1/\sqrt{2})(|0000\rangle + |1111\rangle),$$

$$|01\rangle \rightarrow (1/\sqrt{2})(|1100\rangle + |0011\rangle),$$

$$|10\rangle \rightarrow (1/\sqrt{2})(|1010\rangle + |0101\rangle),$$

$$|11\rangle \rightarrow (1/\sqrt{2})(|0110\rangle + |1001\rangle).$$

Using the codewords corresponding to $|00\rangle$ and $|01\rangle$ basis states, show why this code can detect a single qubit bit flip, but cannot correct it. [3 marks]

(e) For the 4-qubit code in (d), using the codewords corresponding to $|00\rangle$ and $|01\rangle$ basis states, show why this code can detect a single qubit phase flip, but cannot correct it. [2 marks]

(f) For the 4-qubit code in (d), design a circuit that can detect a single qubit bit flip error in the codeword qubits. Use only 1 ancilla qubit to perform error detection. [4 marks]

(g) For the 4-qubit code in (d), design a circuit that can detect a single qubit phase flip error in the codeword qubits. Use only 1 ancilla qubit to perform error detection. [4 marks]

12 Randomised Algorithms

In the weighted vertex cover problem, we are given an undirected graph $G = (V, E)$, $n = |V|$ with a weight function $w : V \to \mathbb{R}^+$. We seek a subset $C \subseteq V$ such that each edge has at least one endpoint in C and $w(C) = \sum_{v \in C} w(v)$ is minimised.

(a) Express the weighted vertex cover problem as an integer program. [3 marks]

Consider the following deterministic greedy algorithm.

Input: $G = (V, E)$, $w : V \to \mathbb{R}^+$ with edges $e_1, e_2, \ldots, e_m$
- $C = \emptyset$
- For $i = 1$ to m do
- Let $e_i = \{u, v\}$
- If $\{u, v\} \cap C = \emptyset$ then
- If $w(u) < w(v)$ then $C \leftarrow C \cup \{u\}$ else $C \leftarrow C \cup \{v\}$

(b) Show that the approximation ratio of this algorithm is unbounded in n. [4 marks]

Consider now the following randomised algorithm.

Input: $G = (V, E)$, $w : V \to \mathbb{R}^+$ with edges $e_1, e_2, \ldots, e_m$
- $C = \emptyset$
- For $i = 1$ to m do
- Let $e_i = \{u, v\}$
- If $\{u, v\} \cap C = \emptyset$ then
- With prob. $\frac{w(v)}{w(u)+w(v)}$ update $C \leftarrow C \cup \{u\}$, otherwise $C \leftarrow C \cup \{v\}$

(c) Assuming that the above is a randomised 2-approximation algorithm, design a new algorithm such that the returned cover C and optimal cover C^* satisfy $\mathbf{P}\,[\, w(C) \leq 3 \cdot w(C^*)\,] \geq 1 - n^{-1}$. [4 marks]

(d) Consider the following simplified input. For any integer $k \geq 0$, the star graph $G_k = (V, E)$ is given by $V = \{v, u_1, u_2, \ldots, u_k\}$ and $E = \{\{v, u_i\} \colon 1 \leq i \leq k\}$; thus v is a "center" node connected to all other nodes. Let C_k be the returned solution of the randomised algorithm on G_k.

(i) Prove that

$$\mathbf{E}\,[\, w(C_k)\,] \leq 2 \cdot w(v).$$

Hint: Find a recursive formula for $\mathbf{E}\,[\, w(C_k)\,]$ and then use induction over k. [5 marks]

(ii) Using the result from (d), prove that the randomised algorithm is a randomised 2-approximation algorithm for any input graph.
Hint: Decompose the edge-set E into star graphs. [4 marks]

13 Types

(*a*) Consider the following OCaml type:

```
type bexp = True | False | Not of bexp | And of bexp * bexp
```

In this question we will look at its encoding in System F.

(*i*) Give a suitable System F type for a Church encoding of the `bexp` type. [2 marks]

(*ii*) Give an implementation of the `True`, `False`, `Not` and `And` constructors for this encoding. [4 marks]

(*iii*) Give the type and encoding of the recursive eliminator named fold for this tree type. [2 marks]

(*iv*) Give reduction rules for fold. [3 marks]

(*v*) For the `And` case, show that your encoding models the reduction rule correctly. [4 marks]

(*b*) (*i*) Using the simply-typed lambda calculus augmented with state and integers, write a function $\mathsf{count} : ((\mathsf{unit} \to \mathsf{unit}) \to \mathsf{unit}) \to (\mathsf{unit} \to \mathsf{unit}) \to \mathsf{int}$, such that $\mathsf{count}\, k\, f$ computes $k\, f$, and returns the number of times k invokes the function f. [3 marks]

(*ii*) In the pure, total simply-typed lambda calculus with integers, characterise the behaviour of a function $h : ((\mathsf{unit} \to \mathsf{unit}) \to \mathsf{unit}) \to (\mathsf{unit} \to \mathsf{unit}) \to \mathsf{int}$ in terms of its arguments k and f. [2 marks]

END OF PAPER