CST2
COMPUTER SCIENCE TRIPOS Part II

Tuesday 4 June 2024 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

Many high-performance microprocessors support multithreading in hardware.

(*a*) In coarse-grained multithreading, threads switch following specific events.

(*i*) What hardware support is required for coarse-grained multithreading? [3 marks]

(*ii*) What hardware can be provided to reduce the cost of thread switching in coarse-grained multithreading and how does it help? [3 marks]

(*b*) In fine-grained multithreading, a new thread is selected to be fetched on each clock cycle.

(*i*) How can fine-grained multithreading reduce the hardware requirements of a simple in-order processor in some circumstances? [3 marks]

(*ii*) What is the impact on performance of fine-grained multithreading and how can it be improved? [3 marks]

(*c*) In simultaneous multithreading, threads co-exist within a core.

(*i*) Describe a scenario where overall performance will improve, and another where it will get worse, with simultaneous multithreading. [4 marks]

(*ii*) What factors need to be considered when deciding whether to duplicate, partition or share a core resource? [4 marks]

2 Bioinformatics

(*a*) You are a data scientist working at a hospital. A former in-patient claims to have been infected with HIV during their previous stay in the hospital. You have access to blood samples of a number of patients who were hospitalised at the same time as the claimant.

(*i*) Describe how you would investigate the claim. [5 marks]

(*ii*) Discuss how to evaluate the robustness of your finding. [5 marks]

(*b*) You are given a table of gene expression data for different patients and controls (healthy people). The rows represent different genes, columns represent a group of patients or controls. Discuss an algorithm to identify the most important genes involved in the disease, its complexity and limitations. [5 marks]

(*c*) A new deep-sea animal species is captured by a British expedition in the Mariana Trench, the deepest trench in the world, and a DNA sample is successfully sequenced. Species in extreme environment have usually adapted by accumulating large numbers of genomic rearrangements and mutations with respect to ancestor species. Discuss DNA coverage and reads length for robust genome assembly in the case the genome of the new species contains large number of repeated genomic regions. [5 marks]

3 Business Studies

After completing your Computer Science degree, you and a small group of friends start a generative AI company to help lecturers produce better-designed slides and course material.

You grow the company successfully and raise a number of funding rounds. Unfortunately you are not yet profitable, and the fund-raising environment changes just as you are about to start your next round. As a result, you decide to push it back by 6 to 9 months.

(*a*) As the Head of Product Development for the company, how would you go about making your development budget last until the next fund-raising round. As part of your answer, consider the implications for your team. [8 marks]

(*b*) Following the delayed fund-raising plan, the board have determined that the CEO has not been "consistently candid" with the board or customers, and needs to be replaced. You have agreed to become the new CEO. How would you manage the transition while minimising the negative impact on the company? [12 marks]

4 Cryptography

(*a*) Consider a cyclic group $(\mathbb{G}, \bullet)$ of order q with generator g.

Briefly explain the difference between the Computational Diffie–Hellman problem and the Decision Diffie–Hellman problem for $\mathbb{G}$, and state how if one of these problems is hard for $\mathbb{G}$, what this implies for the other. [6 marks]

(*b*) While decompiling the executable of an ECDSA implementation with unknown domain parameters, you encounter a prime-number constant of the form

```
0x ffffffff ffffffff fffffffe 0626bd0c 2f33945b 7d67dbcb
```

Based on the structure of its hexadecimal representation, what rôle could this number play? Explain your answer based on how elliptic-curve groups used in cryptography can be constructed. [6 marks]

(*c*) A certification authority C would like to issue certificates that bind a user A's public key PK_A to not just that user's name, but to 10 different personal attribute values $A_0, \ldots, A_9$, e.g. forename, surname, year of birth, birthday, gender, country, postcode, street address, email, portrait photo. User A can then use such a certificate to register with a range of different online services. However, not all attributes are required, or even appropriate, to be revealed to each service: some may only need the email address, whereas others need perhaps only forename, year of birth, gender, and the photo.

User A should, therefore, be able to choose, which subset $S \subset \mathbb{Z}_{10}$ of these 10 attributes they want to reveal each time they present their certificate to a service. One solution would be that C signs for each user 2^{10} different certificates, each including a different subset of attributes. But that would be rather inefficient.

Propose a certificate format, where C generates just one digital signature for each user A, but A then can modify their certificate to remove any subset of the ten attribute values, such that the recipient still can be sure the received attribute values are authentic, while not being able to infer the value of the removed attributes (except with negligible probability in polynomial time). Explain in detail what A receives from the certification authority, and what A provides to a service that only needs a certificate covering attribute subset S. [8 marks]

5 Denotational Semantics

In all parts of this question, you are allowed to use theorems from the course, provided you state them precisely beforehand. You may also extend a proof by (rule) induction from the course with new cases without reproving the ones from the course, again provided you clearly state the proof you are extending.

(*a*) Given domains D_1, D_2, E_1 and E_2, and continuous functions $f_1 \colon D_1 \to E_1$ and $f_2 \colon D_2 \to E_2$, show that

$$\begin{array}{rccl} f_1 \times f_2 \colon & D_1 \times D_2 & \to & E_1 \times E_2 \\ & (d_1, d_2) & \mapsto & (f_1(d_1), f_2(d_2)) \end{array}$$

is continuous. [6 marks]

We wish to extend PCF with the product type $\tau_1 * \tau_2$, by adding the new terms `fst`, `snd` and `pair` to the language, such that

$$\frac{\Gamma \vdash t \colon \tau_1 * \tau_2}{\Gamma \vdash \texttt{fst}(t) \colon \tau_1} \qquad \frac{\Gamma \vdash t \colon \tau_1 * \tau_2}{\Gamma \vdash \texttt{snd}(t) \colon \tau_2} \qquad \frac{\Gamma \vdash t_1 \colon \tau_1 \quad \Gamma \vdash t_2 \colon \tau_2}{\Gamma \vdash \texttt{pair}(t_1, t_2) \colon \tau_1 * \tau_2}$$

and with the following operational semantics:

$$\frac{t \Downarrow_{\tau_1 * \tau_2} \texttt{pair}(v_1, v_2)}{\texttt{fst}(t) \Downarrow_{\tau_1} v_1} \qquad \frac{t \Downarrow_{\tau_1 * \tau_2} \texttt{pair}(v_1, v_2)}{\texttt{snd}(t) \Downarrow_{\tau_2} v_2} \qquad \frac{t_1 \Downarrow_{\tau_1} v_1 \quad t_2 \Downarrow_{\tau_2} v_2}{\texttt{pair}(t_1, t_2) \Downarrow_{\tau_1 * \tau_2} \texttt{pair}(v_1, v_2)}$$

(*b*) Give a denotational semantics for the product type $[\![\tau_1 * \tau_2]\!]$ in terms of $[\![\tau_1]\!]$ and $[\![\tau_2]\!]$. [2 marks]

(*c*) Give a denotational semantics for `fst`, `snd` and `pair`, extending the semantics of PCF from the lectures, and justify why this semantics is well-defined according to the typing rules given above. [6 marks]

(*d*) Recall what it means for denotational semantics to be sound. Show that the semantics you have just given is sound. [6 marks]

6 Hoare Logic and Model Checking

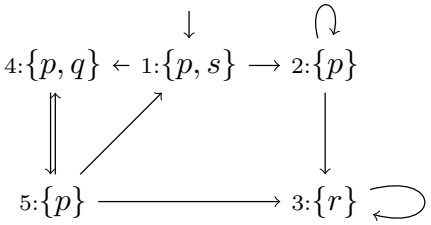
Consider the temporal logic CTL over atomic propositions $p \in \mathit{AP}$:
$\psi \in \mathsf{StateProp} ::= \bot \mid \top \mid \neg\psi \mid \psi_1 \wedge \psi_2 \mid \psi_1 \vee \psi_2 \mid \psi_1 \rightarrow \psi_2 \mid p \mid \mathsf{A}\ \phi \mid \mathsf{E}\ \phi$,
$\phi \in \mathsf{PathProp}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

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

(i) If a state satisfying p cannot be reached, then q always holds. [3 marks]

(ii) From all reachable states, there is some path along which p holds, until it reaches a state from which no possible next state satisfies q. [3 marks]

(b) Consider a temporal model M over atomic propositions $\mathit{AP} = \{p, q, r, s\}$, with states $\{1, 2, 3, 4, 5\}$, initial state 1, and transitions and state labelling as shown in the diagram (e.g. in state 1, atomic propositions p and s 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) $\mathsf{A}((q \wedge s)\mathsf{U}(\mathsf{EF}r))$ [2 marks]

(ii) $\mathsf{EG}(p \wedge \mathsf{AX}p)$ [3 marks]

(c) (i) Informally explain the difference in the properties that can be expressed by LTL and CTL. [3 marks]

(ii) Consider the LTL formula $\phi = p\mathsf{U}(\mathsf{X}q)$ and CTL formula $\psi = \mathsf{A}(p\mathsf{U}(\mathsf{AX}q))$, both over atomic propositions $\mathit{AP} = \{p, q\}$. Formally define a temporal model over AP that shows that ϕ and ψ are not equivalent. Explain why your temporal model satisfies one of the formulae but not the other. [6 marks]

7 Information Theory

(a) Draw a diagram that relates the mutual information between two random variables to their entropies, conditional entropies and joint entropy. [2 marks]

(b) In the analysis of continuous signals, explain why we often constrain the variance of the signal. What input distribution gives the maximum entropy under this constraint? [3 marks]

(c) What is the Gaussian channel? Why is it particularly relevant in the analysis of real world communications systems? [3 marks]

(d) Consider a Gaussian channel with input, output and noise represented by random variables X, Y and Z, such that Y = X + Z. State with justification, but without a detailed proof, the probability distribution of X that achieves the capacity. Derive an expression for this capacity.

[*Note:* You may use the result that the entropy of a normally distributed random variable $X \sim N(\mu, \sigma^2)$ is $\frac{1}{2}\log(2\pi e\sigma^2)$ without proof.] [7 marks]

(e) The Nyquist sampling theorem says that a signal with maximum frequency f must be sampled at no less than least $2f$ to allow reconstruction. Use this, together with your answer to (d), to derive the capacity of a Gaussian channel where the noise has bandwidth limited to B. [2 marks]

(f) Use your answer to (e) to explain how an ultra-wideband (UWB) communications system (with bandwidths of multiple GHz) can avoid interference with other non-UWB users of the same part of the radio spectrum. [3 marks]

8 Machine Learning and Bayesian Inference

(a) Consider the **support vector machine** (SVM) for inputs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n, (\mathbf{x}, y) \in \mathcal{X} \times \{0, 1\}$, $\mathcal{X} \subset \mathbb{R}^d$. Let $(\mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R}$ denote the parameters which define the maximum-margin hyperplane returned by the SVM.

(i) The SVM classification function is given by the maximum-margin hyperplane:

$$f(\mathbf{x}) = \text{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle + b).$$

Express $\mathbf{w}$ in terms of the dual variables $\{\lambda_i\}_{i=1}^n$ associated with the margin violation constraints. Hence rewrite $f(\mathbf{x})$ in terms of λ_i and interpret how the hypothesis function classifies an unseen point $\mathbf{x}^*$. [4 marks]

(ii) What property of the hypothesis function allows the extension of the SVM to define a nonlinear decision boundary in the feature space $\mathcal{X}$? State the nonlinear extension of $f(\mathbf{x})$ and explain how this may improve classification performance. [4 marks]

(iii) Consider the SVM primal objective, where $C > 0$ is a constant:

$$\underset{\mathbf{w}, \boldsymbol{\xi}}{\text{argmin}} \left(\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \right).$$

Let $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ be the Gram matrix evaluated on training points, where $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a valid kernel function. Assume we have subsumed the bias b into the kernel. Write the kernelised SVM objective in terms of $K, \boldsymbol{\alpha}$ and $\boldsymbol{\xi}$, where $\boldsymbol{\alpha}_i = \lambda_i y_i$. [3 marks]

(iv) Rewrite the SVM objective in terms of the hypothesis function applied to each example, $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^T$. [*Hint:* Relate $\mathbf{f}$ and $\boldsymbol{\alpha}$.] [3 marks]

(b) Consider modelling the data previously given as an underlying function contaminated with additive Gaussian noise, $y_i = f(\mathbf{x}_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Model the function f as a Gaussian process with zero mean, and covariance function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, i.e. $f \sim \mathsf{GP}(0, \kappa)$.
[*Hint:* Recall for a d-dimensional normally distributed random variable $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $p(\mathbf{z}) = \left((2\pi)^d \det \Sigma\right)^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{z} - \boldsymbol{\mu})\right)$.]

(i) Let $\mathbf{f}, \mathbf{y}, X$ denote the collection of function values, training labels and feature vectors, respectively. Write down the log-prior, $\log p(\mathbf{f}|X)$ in terms of the Gram matrix $\mathbf{K}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$. [2 marks]

(ii) Find the posterior over $\mathbf{f}$, $p(\mathbf{f}|X, \mathbf{y})$. Neglect the normalisation and compute the un-normalised log-posterior, neglecting constant terms. Compare against the SVM objective found in Part (a)(iv). [4 marks]

9 Optimising Compilers

Compilers for functional languages sometimes perform strictness analysis to pass parameters by value rather than by name.

(*a*) Define the concept of *strictness* and how it differs from *neededness*. [3 marks]

(*b*) Write functions with two parameters for each of the following cases.

(*i*) A function that is strict in both parameters but only the first is needed. [2 marks]

(*ii*) A function that is strict in its first parameter and only the first is needed. [2 marks]

(*iii*) A function that is strict in neither parameter and neither is needed. [1 mark]

(*c*) Perform strictness analysis on the following program to obtain its strictness function. In which parameter(s) is `f` strict?

```
f(a, b, c)
  = if a<1 then b elif a<2 then c else f(a-c, b, c);
```

You may use the following built-in strictness functions.

$$\begin{aligned} 1^\sharp &= 1 \\ 2^\sharp &= 1 \\ lt^\sharp(x,y) &= x \wedge y \\ sub^\sharp(x,y) &= x \wedge y \\ cond^\sharp(p,x,y) &= p \wedge (x \vee y) \end{aligned}$$

[6 marks]

(*d*) After strictness optimisation, some parameters remain passed by name, yet you wish to evaluate these as early as possible within the function whilst maintaining strictness properties. Describe an analysis to do this, and explain the results on the program in part (*c*). [6 marks]

10 Principles of Communications

(*a*) The Internet is a shared resource. Users compete to send traffic, but need to cooperate to conserve resource. However user traffic has two fundamentally different utility curves, being elastic, or inelastic.

In designing resource sharing schemes, two different fairness goals have been defined: proportional fairness versus max-min fairness.

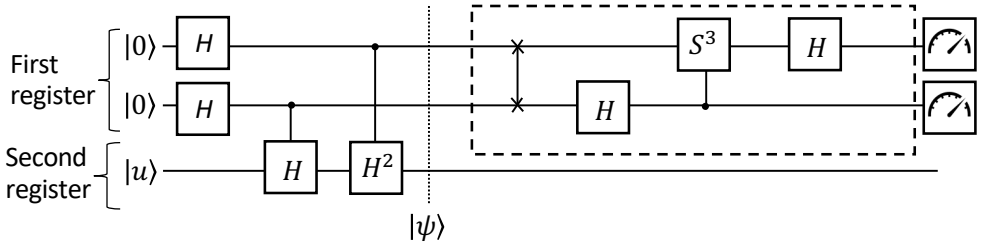
How do these goals reflect the traffic requirements of the two different utility curves? [10 marks]

(*b*) Routers can support different types of schedulers to provide fairness and isolation between traffic flowing between different sources and destinations.

You have read about fair queueing, and hear someone has proposed a simpler scheme which could remove the requirement for per-flow state in the scheduler. The proposal is to use random scheduler. Would that be fair? What about isolation?

What are the general considerations about traffic destined to be handled by such a scheduler, for it to work reasonably well? [10 marks]

11 Quantum Computing



(*a*) The figure shows the circuit for quantum phase estimation of a Hadamard gate. What is the function of the sub-circuit shown in the box marked with the dashed line, and to how many bits of precision is the estimate of the phase given? [2 marks]

(*b*) The Hadamard gate has matrix $\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$. What are its eigenvectors and corresponding eigenvalues? Express each eigenvector as a quantum state (that is, as superposition of computational basis states). [5 marks]

(*c*) Simplify the circuit in the figure such that when the initial state of the first register is $|00\rangle$ as specified, the top wire only involves a swap gate and a measurement. [6 marks]

(*d*) Quantum phase estimation is performed using the circuit given in the figure with $|u\rangle = a|0\rangle + b|1\rangle$. Express the three-qubit state $|\psi\rangle$ in terms of a and b. Verify that if $|u\rangle$ is a correctly normalised quantum state then so is $|\psi\rangle$. [7 marks]

12 Randomised Algorithms

Given an undirected graph $G = (V, E)$, an **independent set** is a subset $I \subseteq V$ such that for any two vertices $u \in I, v \in I$, there is no edge $\{u, v\} \in E(G)$. Let $\alpha(G)$ denote the size of the **largest** independent set in G.

(a) Consider the following randomised algorithm for computing an independent set, which takes as input an undirected graph $G = (V, E)$ and a fixed parameter $p \in [0, 1]$:

Step 1: Starting with an empty set S, add each vertex from $V(G)$ to S independently with probability p.

Step 2: Go through all edges $e = \{u, v\} \in E(G)$, and for any edge e which had both vertices in S after Step 1, remove u **or** v from S.

(i) Justify briefly why the output S of this algorithm is an independent set. [2 marks]

(ii) Is the output S necessarily maximal, i.e., it is not possible to add any vertex $u \in V$ to S and obtain a larger independent set? Justify your answer. [3 marks]

(iii) Prove that the expected size of the output S after the second step of the algorithm is at least $p \cdot |V| - p^2 \cdot |E|$. [4 marks]

(iv) How would you choose p in order to maximise the expected size of S, as computed in (a)(iii)? [4 marks]

(v) What does your answer in (a)(iv) imply for $\alpha(G)$? Justify your answer. [3 marks]

(b) Formulate the problem of finding the largest independent set as an Integer Program (**I**), and describe the Linear Programming Relaxation (**L**). [4 marks]

13 Types

(*a*) Derive the following entailments with the natural deduction system for classical logic.

(*i*) Show that $A \vee B; A \vdash B\,\mathsf{true}$. [5 marks]

(*ii*) Show that $A \vee B, \neg A; \cdot \vdash B\,\mathsf{true}$. [7 marks]

(*b*) Consider System F extended with existential types, products, and a natural number type.

(*i*) Give a Church encoding for an optional natural number type (corresponding to `nat option` in OCaml). [2 marks]

(*ii*) Give an existential type corresponding to an abstract type of optional naturals, with constructors for `Some` and `None`, as well as a case analysis operation. It should correspond to the following OCaml module signature:

```
module type ONAT = sig
  type t
  val none : t
  val some : nat -> t
  val case : t -> 'a -> (nat -> 'a) -> 'a
end
```

[3 marks]

(*iii*) Give an implementation of this existential type. [3 marks]

END OF PAPER