CST2
COMPUTER SCIENCE TRIPOS Part II

Tuesday 6 June 2023 14:00 to 17:00

COMPUTER SCIENCE Paper 9

*Answer **five** questions.*

*Submit each question answer in a **separate** PDF. As the file name, use your candidate number, paper and question number (e.g., `1234A-p9-q6.pdf`). Also write your candidate number, paper and question number at the start of each PDF.*

You must follow the official form and conduct instructions for this online examination

1 Advanced Computer Architecture

(*a*) Describe why it is often important to try to limit off-chip memory accesses when designing a high-performance domain-specific accelerator that is part of a System-on-a-Chip (SoC)? [4 marks]

(*b*) What techniques or strategies might be employed to reduce off-chip memory accesses when designing a domain-specific accelerator? [8 marks]

(*c*) Imagine a domain-specific accelerator with numerous individual compute units. Each compute unit is programmable, has specialised functional units and produces requests to access memory. The compute units share a L1 cache.

In order to help improve the effective bandwidth to the shared L1 cache, memory requests that access the same address may be coalesced (or merged). Experiments confirm that there is scope to coalesce requests both from a single unit and from different units.

You are asked to design a hardware scheme for coalescing load requests only, store requests are not coalesced. The compute units generate loads and stores in program order. For all requests from a single compute unit, the coalescing unit must ensure that stores are not reordered with respect to loads or other stores from the same compute unit. The accelerator provides no guarantees about the way in which requests from different compute units may be reordered.

The design should be efficient and scalable, i.e. make good use of any on-chip memory required and make some attempt to minimise the number of address comparisons required. It should be possible to coalesce requests that are made from different compute units. There is no need to discuss how data is returned from the L1 cache to the compute units.

Describe your design, list any assumptions and justify your decision decisions. [8 marks]

2 Bioinformatics

(a) What is the output of the UPGMA algorithm given the distance matrix of the species a, b, c, d, e below?

$$\begin{pmatrix} & a & b & c & d & e \\ a & 0 & 16 & 21 & 31 & 20 \\ b & & 0 & 30 & 234 & 21 \\ c & & & 0 & 28 & 38 \\ d & & & & 0 & 42 \\ e & & & & & 0 \end{pmatrix}$$

[4 marks]

(b) Compute the neighbour joining phylogeny from the four species (s1,s2,s3,s4) DNA sequences.

s1: `GATAA`
s2: `GATAC`
s3: `CTTTC`
s4: `CTGGG`

[4 marks]

(c) Compute the Burrows-Wheeler transform on the string `PARALLELISM`.

[4 marks]

(d) Discuss the concept of modularity in the Louvain algorithm. [4 marks]

(e) Discuss the requirement of well-stirred chemical solution for the Gillespie Algorithm. [4 marks]

3 Business Studies

A PhD student has made an AI-powered chat system available on the Internet, through a technology preview on their departmental webpage. The student has attracted millions of users, and has exhausted their PhD-funded server resource. They have been asked by the department to remove the preview from the university's servers and find a way to fund it themself.

(*a*) Referencing the customer adoption curve, how should the student go about analysing early product use to inform their development plan for the product? [5 marks]

(*b*) Outline a plan to enable the student to build a business around a dialogue-based AI chatbot, including explicit risk mitigation strategies for the 5 major risks to the product's success. [15 marks]

4 Cryptography

A building access-control smartcard uses the following authentication protocol. Let G be a generator for an elliptic-curve based cyclic group $(E(\mathbb{Z}_p, a, b), +)$ of order q. The card stores in its non-volatile memory a secret key $u \in \mathbb{Z}_q^*$ and a public card identifier $U := uG$. The door does not know u. Curve parameters were chosen such that determining u from curve point U is computationally infeasible.

When the user holds the contactless card in front of the door reader, the card picks a number $c \in \mathbb{Z}_q^*$ and the door picks a number $d \in \mathbb{Z}_q^*$, both uniformly at random. The card calculates the coordinates of elliptic-curve point $C := cG$.

They then exchange the following three messages:

card		**door**
read u, U from memory		
pick $c \in \mathbb{Z}_q^*$, $C := cG$	① $\xrightarrow{(U, C)}$	
	② $\xleftarrow{d}$	pick $d \in \mathbb{Z}_q^*$
$r := (c + du) \bmod q$	③ $\xrightarrow{r}$	

(a) What checks should the door perform on the received values U, C, r to verify that the card identified by U really is in possession of u? [4 marks]

(b) How many bits will be required to encode the values exchanged in these three messages in order to achieve a security level similar to the use of a 128-bit key in a symmetric MAC? [4 marks]

(c) Your colleague is concerned that the calculation of $C := cG$ in the card slows down the authentication process too much, and therefore proposes to postpone transmission of C to the third message, i.e. to change the three protocol messages from previously (U, C), d, r to now U, d, (C, r). Would this affect security? [5 marks]

(d) Due to supply-chain issues, the hardware manufacturer no longer can make door readers that send data to the card. Modify the original protocol such that only the card sends data to the door. The card maintains a counter m for how often it has been used, and the door remembers the highest value of m it has previously seen and will only open again when presented with a new value m higher than any seen before. Instead of receiving d in message ②, let the card calculate d in a way such that the card provides a digital signature of m, and sends (d, m) to the card. Keep message ③ the same. [5 marks]

(e) Which value appearing in the original protocol no longer has to be transmitted by the unidirectional variant from Part (d), and why? [2 marks]

5 Denotational Semantics

(a) Consider the following definitions:

- For $L \in \mathrm{PCF}_{nat}$ and $k \in \mathbb{N}$, $L \Vdash_0 k$ if, and only if, $L \Downarrow_{nat} \mathbf{succ}^k(\mathbf{0})$.
- For $M \in \mathrm{PCF}_{nat \to nat}$ and $f : \mathbb{N} \to \mathbb{N}$, $M \Vdash_1 f$ if, and only if, for all $i \in \mathbb{N}$, $M\ \mathbf{succ}^i(\mathbf{0}) \Downarrow_{nat} \mathbf{succ}^{f(i)}(\mathbf{0})$.
- For $N \in \mathrm{PCF}_{nat \to nat \to nat}$ and $g : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$, $N \Vdash_2 g$ if, and only if, for all $i, j \in \mathbb{N}$, $N\ \ \mathbf{succ}^i(\mathbf{0})\ \ \mathbf{succ}^j(\mathbf{0}) \Downarrow_{nat} \mathbf{succ}^{g(i,j)}(\mathbf{0})$.

(i) Prove that $N \Vdash_2 g$ and $L \Vdash_0 k$ imply $N\,L \Vdash_1 \lambda\, x \in \mathbb{N}.\, g(k, x)$. [6 marks]

(ii) Prove that there are $N \in \mathrm{PCF}_{nat \to nat \to nat}$ and a bijection $g : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ such that $N \Vdash_2 g$. You may use standard results provided that you state them clearly. [6 marks]

(b) (i) Say whether or not the following statement holds:

For all PCF types τ and all closed PCF terms M of type $\tau \to \tau \to \tau$, the closed PCF terms $\mathbf{fix}\big(\mathbf{fn}\ x : \tau.\, \mathbf{fix}(\mathbf{fn}\ y : \tau.\, M\ y\ x)\big)$ and $\mathbf{fix}(\mathbf{fn}\ z : \tau.\, M\ z\ z)$ of type τ are contextually equivalent. [2 marks]

(ii) Either prove or disprove the above statement. [6 marks]

6 Hoare Logic and Model Checking

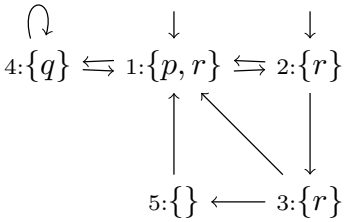
Consider the temporal logic CTL over atomic propositions $p \in \mathit{AP}$:
$\psi \in \mathsf{StateProp} ::= \bot \mid \top \mid \neg\psi \mid \psi_1 \wedge \psi_2 \mid \psi_1 \vee \psi_2 \mid \psi_1 \rightarrow \psi_2 \mid p \mid \mathsf{A}\ \phi \mid \mathsf{E}\ \phi$,
$\phi \in \mathsf{PathProp}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

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

(i) There exists a path such that at some point p will always hold. [2 marks]

(ii) There exists a path such that at some point q holds, and from any state along the path until then, a state satisfying p can be reached. [3 marks]

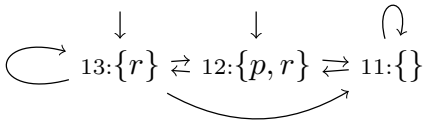
(b) Consider a temporal model M over atomic propositions $\mathit{AP} = \{p, q, r\}$, with states $\{1, 2, 3, 4, 5\}$, initial states 1 and 2, and transitions and state labelling as shown in the diagram (e.g. in state 1, atomic propositions p and r hold). Informally describe the meaning of each of the following CTL formulae over AP and explain why they hold in the model or give a counterexample if they do not.



(i) $\mathsf{A}(r\ \mathsf{U}\ (\mathsf{EX}q))$ [2 marks]

(ii) $(\mathsf{AF}p) \wedge (\mathsf{AGEF}q)$ [3 marks]

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



(i) Show that M' simulates M: define a relation R and show $M \preccurlyeq^R M'$. [6 marks]

(ii) Is your relation R a bi-simulation? Explain why or why not. [4 marks]

7 Information Theory

You are tasked with compressing a continuous textual data stream into a stream of binary codewords. You know the stream's alphabet contains three letters ('A', 'B', 'C') and three numbers ('1', '2', '3'), and that the stream alternates between letter and number. You have been provided with a sample of 100 consecutive characters and no more as follows:

```
A1A3A1C2B2
A3B2A2B2C2
C3C3C3C3A1
B1B2B1C1C1
A1C3A1A2A1
A1A2A1A2B3
A2B3A1A3B3
C2A3A3C2A1
C2C2A3A3C3
C2A3A3B2A3
```

(*a*) For each of the following source models find an encoding using Huffman codes and compare the average encoded character length to the entropy in bits per character.

(*i*) A pure character source, ignoring the alternating nature of the characters. [3 marks]

(*ii*) A mixture of two distinct sources, one for letters and one for numbers. [4 marks]

(*iii*) A stream of two-character symbols ('A1', 'A2', etc.). [5 marks]

(*b*) Explain conceptually the trend in entropy values you found in part (*a*). Which do you think is closer to the true entropy and why? [4 marks]

(*c*) Discuss whether it would be advantageous to model the stream as a stream of four-character symbols and apply Huffman coding. [2 marks]

(*d*) Give two advantages to using arithmetic coding instead of Huffman coding for this problem. [2 marks]

8 Machine Learning and Bayesian Inference

You are faced with the following simple inference problem. Examples are pairs (x, y), where both x and y are real numbers. You suspect that there is a relationship $y = \theta x$ underlying the data, and that the y values are subject to additive Gaussian noise with mean 0 and variance σ^2. You wish to infer θ from examples (x_i, y_i) where $i = 1, \ldots, m$. You may consider x values to be fixed.

(a) Write down an expression for the density $p(Y|\theta; x)$. You may use the standard expression for the Gaussian density $N(\mu, \sigma^2)$

$$p(Z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(Z-\mu)^2\right).$$

[1 mark]

(b) Write down an expression for the likelihood of the parameter θ, given the m training examples described. State any assumptions you make. [3 marks]

(c) Assuming that the parameter θ has a Gaussian prior with mean 0 and variance σ_1^2, write down an expression for the posterior density $p(\theta|\mathbf{y}; \mathbf{x})$ of θ. You may leave this expression unnormalized. State any assumptions you make.
[3 marks]

(d) Explain how the calculation in Part (c) might be extended to obtain the *evidence*, and how this might be used to estimate the values of σ and σ_1 from the training data. State any assumptions needed. You need not perform the actual calculation. [3 marks]

(e) Find an expression for the *predictive density* $p(Y|x, \mathbf{x}, \mathbf{y})$. You need not normalize the expression. You may use the identity

$$\int_{\mathbb{R}^p} \exp\left(-\frac{1}{2}\left(\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{b} + c\right)\right) = (2\pi)^{p/2}|\mathbf{A}|^{-1/2} \exp\left(-\frac{1}{2}\left(c - \frac{\mathbf{b}^t\mathbf{A}\mathbf{b}}{4}\right)\right).$$

[7 marks]

(f) What further steps are necessary should you wish to extend your result from Part (e) to obtain actual predictions for Y along with a measure of certainty in the prediction? [3 marks]

9 Optimising Compilers

Compilers use intermediate representations (IRs) when optimising code.

(*a*) For each of the IRs below, describe its merits for performing dead code elimination, providing a diagram or pseudo-code to illustrate your answer.

(*i*) An abstract syntax tree. [4 marks]

(*ii*) A three-address code. [4 marks]

(*iii*) A stack-based IR. [4 marks]

(*b*) A decompiler is given the following IR code for a function, where registers `r0` and `r1` contain the function arguments and `r0` contains the returned value:

```
foo: mov   r2, #0        // r2 = 0
     cmple r0, #0, end   // Branch to end if r0 <= 0
top: shl   r0, r0, #2    // r0 = r0 << 2
     ldr   r0, [r1, r0]  // r0 = MEM[r1 + r0]
     cmpne r0, #5, chk   // Branch to chk if r0 != 5
inc: add   r2, r2, #1    // r2 = r2 + 1
chk: cmpgt r0, #0, top   // Branch to top if r0 > 0
end: mov   r0, r2        // r0 = r2
     ret                 // Return
```

(*i*) Draw the dominance tree for this code. [2 marks]

(*ii*) Reconstruct a representative source code for this IR code assuming all types are integers or pointers. [6 marks]

10 Principles of Communications

(*a*) One simple equation for mean throughput experienced by a TCP connection is proportional to

$$\frac{MSS}{RTT \cdot \sqrt{p}}$$

where MSS is the maximum segment size, or in other words, the data packet size; RTT is the round trip time; and p is the packet loss probability.

Explain this equation, starting from the basic AIMD behaviour of TCP's congestion control and avoidance window adjustment algorithm. Assume that a TCP flow is unidirectional and all data packets are maximum size, and that packet losses only happen due to congestion. In your answer, please explain any other assumptions, for example concerning the round trip time and link capacities on the path from sender to recipient. [10 marks]

(*b*) A clever person decides to mitigate the $1/RTT$ "unfairness" in the TCP throughput by building a scheduler that uses weighted round robin, and assigns weights to TCP flows, in *inverse proportion* to their RTT.

Give reasons why this might not be an appropriate solution. Also give reasons why this might not be easy to implement. [10 marks]

11 Quantum Computing

(*a*) (*i*) State and prove the no-cloning principle. [2 marks]

(*ii*) Using standard quantum gates, construct a 2-qubit unitary circuit U that copies the orthogonal states $|+\rangle$ and $|-\rangle$. That is, U should have the properties $U(|+\rangle|0\rangle) = |+\rangle|+\rangle$ and $U(|-\rangle|0\rangle) = |-\rangle|-\rangle$. [3 marks]

(*b*) An experimenter has prepared an initial 2-qubit state, and plans to measure the first qubit in the computational basis, and then measure the second qubit. For each of the following statements, identify the initial states that make it true.

(*i*) If the second qubit is measured in the computational basis, the outcome of the second measurement can always be predicted with certainty, once the outcome of the first measurement is known.

(*ii*) If the second qubit is measured in the computational basis, the outcome of the second measurement can sometimes be predicted with certainty, depending on the outcome of the first measurement.

(*iii*) Regardless of the choice of basis for the measurement of the second qubit, the probability of obtaining each possible outcome for the second measurement is independent of the outcome of the first measurement.

[5 marks]

(*c*) The experimenter plans to adapt the scheme from *(b)*, so that the basis for the second measurement can be chosen in a way that depends on the first measurement outcome. Is it always possible to choose a measurement basis such that both outcomes for the second measurement have a 50% probability? Justify your answer. [2 marks]

(*d*) Quantum phase estimation is to be performed for this 2-qubit unitary:

$$U = T \otimes H = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 \\ 0 & 0 & 1/2 + i/2 & 1/2 + i/2 \\ 0 & 0 & 1/2 + i/2 & -1/2 - i/2 \end{bmatrix}$$

(*i*) Derive the eigenvectors, eigenvalues and eigenvalue phases of U, showing your working. [4 marks]

(*ii*) If the second register is initialised in the state $|00\rangle$ what are the possible outcomes of running quantum phase estimation, and what is the probability with which each occurs? How many bits of precision are required for QPE to correctly estimate the phase for any initial state? [4 marks]

12 Randomised Algorithms

Consider the following problem (**P**). We are given a *directed* graph $G = (V, E)$ with non-negative edge-weights $w_{i,j} \geq 0$ for each edge $(i, j) \in E$. The task is to partition V into two sets V_1 and $V_2 = V \setminus V_1$ so as to maximize the total weight of edges going from V_1 and V_2, i.e.,

$$\text{maximise} \sum_{(i,j)\in E:\ i\in V_1, j\in V_2} w_{i,j}.$$

(a) Design a randomised approximation algorithm for (**P**) with running time $O(V)$ and analyse its approximation ratio. [5 marks]

(b) In this part of the question, we additionally want that $|V_1| = |V_2| = n/2$ (we assume for simplicity that $n = |V|$ is an even integer). Adjust the algorithm and analysis from (a). [4 marks]

(c) Consider the following integer program called (**I**):

$$\begin{aligned}
\text{maximise} \quad & \sum_{(i,j)\in E} w_{i,j} z_{i,j} & & \\
\text{subject to} \quad & z_{i,j} \leq x_i & & \text{for each } (i,j) \in E \\
& z_{i,j} \leq 1 - x_j & & \text{for each } (i,j) \in E \\
& x_i \in \{0,1\} & & \text{for } i \in V \\
& z_{i,j} \in [0,1] & & \text{for each } (i,j) \in E
\end{aligned}$$

(i) Prove that this integer program solves the problem (**P**). [4 marks]

(ii) Consider the following randomised algorithm for (**P**). Let $(\overline{z}, \overline{x})$ be a solution to a linear program, which is identical to (**I**) but with $x_i \in \{0,1\}$ replaced by $x_i \in [0,1]$. We then put each vertex i into V_1 with probability $1/4 + \overline{x}_i/2$, independently.

(A) Explain briefly why this algorithm can be implemented in polynomial time. [1 mark]

(B) Prove that the approximation ratio of this algorithm is 2. [6 marks]

13 Types

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

(*i*) Show $\neg(A \vee B); \cdot \vdash \neg A$ true. [5 marks]

(*ii*) Show $\cdot; \neg A \vee \neg B \vdash A$ true. [5 marks]

(*b*) (*i*) Using $fold : \forall a.\, a \to (X \to a \to a) \to \mathsf{List}_X \to a$, $cons : X \to \mathsf{List}_X \to \mathsf{List}_X$ and $nil : \mathsf{List}_X$, write a System F function which appends two lists. [1 mark]

(*ii*) Give an OCaml data structure corresponding to the following Church encoding:

$$\forall a.\, a \to (a \to X \to a \to a) \to a$$

[2 marks]

(*iii*) Give a System F term which converts an element t of the type in part (*ii*) of this question into a list with the same elements. [3 marks]

(*c*) Consider the following two Agda proofs:

$$\begin{array}{lcl} \mathsf{unitl} & : & \forall x \to 0 + x \equiv x \\ \mathsf{unitl}\ x & = & \mathsf{refl}(x) \end{array} \qquad \begin{array}{lcl} \mathsf{unitr} & : & \forall x \to x + 0 \equiv x \\ \mathsf{unitr}\ 0 & = & \mathsf{refl}(0) \\ \mathsf{unitr}\ (\mathsf{s}\ n) & = & \mathsf{cong}\ \mathsf{s}\ (\mathsf{unitr}\ n) \end{array}$$

Explain why they are different. [4 marks]

END OF PAPER