CST2
COMPUTER SCIENCE TRIPOS Part II

Wednesday 8 June 2022 11:00 to 14:00 BST

COMPUTER SCIENCE Paper 8

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-p8-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*) A simple snoopy coherence protocol requires that bus transactions are broadcast and observed by all processors in the same order. How could a ring interconnect be used to support this protocol? [4 marks]

(*b*) Imagine a multicore processor with private L2 caches and an inclusive L3 or Last-Level Cache (LLC). We wish to reduce the miss rate of the L2 caches so replace the LLC with a smaller non-inclusive cache allowing for cache chip area to be redistributed to create larger L2 caches. We then discover that the miss rate of the smaller LLC changes very little, why might this be? [3 marks]

(*c*) In the case of a directory-based cache coherence protocol, why might we want to retain an inclusive directory even if our LLC is non-inclusive? [3 marks]

(*d*) It is suggested that directory-based cache coherence can scale with the aid of a hierarchy of on-chip caches. For example, we could group 64 cores into 8 clusters of 8 cores each. Each processor has its own private cache and each cluster has its own shared inclusive "cluster" cache. The chip also contains a shared inclusive Last-Level Cache (LLC). We assume sharers are tracked precisely. Describe how the cache hierarchy can be exploited to reduce the storage cost of tracking sharers. [4 marks]

(*e*) Imagine a System-on-a-Chip (SoC) that consists of multiple cores and a Domain-Specific Accelerator (DSA). The DSA could be given its own cache and be kept fully coherent with the other on-chip caches. Alternatively, the accelerator may be non-coherent and instead DMA data directly from main memory into a local scratchpad. When might each approach be preferable? [6 marks]

2 Bioinformatics

(*a*) What is the purpose of the boostrap approach in general and how can it be applied to phylogenetic trees? Using at least one numeric example, discuss how to interpret bootstrap values. [4 marks]

(*b*) What are the reasons for using progressive alignment in a multi-sequence alignment problem? Give the complexity of the various stages of the procedure and the overall complexity. [4 marks]

(*c*) Define the role of a scoring matrix in a matching algorithm and explain how it should be designed. [3 marks]

(*d*) Sketch the suffix tree for the genome GCTATA$. Give the time and space complexities of using a suffix tree for genome sequence assembly. Comment on finding repeated sequences. [5 marks]

(*e*) We often use Hidden Markov Models to predict genes, exons or introns. Outline how a Hidden Markov Model can be used as a binary classifier in such an application. What metrics can be used to evaluate its performance? [4 marks]

3 Cryptography

(*a*) Let $\Pi = (\mathsf{Gen}, \mathsf{Enc}, \mathsf{Dec})$ be a public-key encryption scheme that offers CCA security. Explain the concept of *forward secrecy*, why it might be useful, and why Π does not offer it. [3 marks]

(*b*) Explain how the Diffie–Hellman key exchange works, and the assumptions under which it is secure. [3 marks]

(*c*) You and your colleague are asked to design a payments system based on an authenticated symmetric encryption scheme $(\mathsf{Enc}, \mathsf{Dec})$, a digital signature scheme $(\mathsf{Gen}, \mathsf{Sign}, \mathsf{Vrfy})$, a Diffie–Hellman group with generator g, and a key derivation function KDF. The requirements are as follows:

- Let B be a bank, and let Alice (A) be a customer of B. Say A has a digital token T (which we take to be an arbitrary bit string) that is worth money. A can deposit that money in her account by securely sending T to B.
- You may assume that the bank knows the public keys of all of its customers, and that each customer knows the public key of the bank.
- As the token T is sent over the network, it must be kept confidential from active attackers. Moreover, the protocol must provide forward secrecy.

Let $(PK_A, SK_A) \leftarrow \mathsf{Gen}$ be Alice's signature keypair, and $(PK_B, SK_B) \leftarrow \mathsf{Gen}$ be the bank's keypair. Your colleague proposes using the following scheme:

$B \rightarrow A: \ (g^x, \mathsf{Sign}_{SK_B}(g^x))$

A receives (g^x, S) and checks whether $\mathsf{Vrfy}_{PK_B}(g^x, S) = 1$.
If this succeeds, A calculates $K = \mathsf{KDF}((g^x)^y)$ and sends:

$A \rightarrow B: \ (g^y, \mathsf{Sign}_{SK_A}(g^y), A, \mathsf{Enc}_K(T))$

B receives (g^y, S, N, C) where N is a customer name, looks up N's public key PK_N, and checks that $\mathsf{Vrfy}_{PK_N}(g^y, S) = 1$; if successful, B decrypts $\mathsf{Dec}_{\mathsf{KDF}(g^{xy})}(C) = T$ and credits it to the account belonging to N.

Let Mallory (M) be an active adversary who is also a customer of the bank. Show that your colleague's scheme is not secure: when Alice wants to deposit a token T in her account, M can cause his account to be credited instead. [7 marks]

(*d*) Suggest an alternative protocol that meets the requirements in part (*c*) while avoiding the problems in your colleague's scheme, and briefly justify your design. [7 marks]

4 Denotational Semantics

(*a*) For posets D and E, and for monotone functions $f : D \to E$ and $g : E \to D$, we write $f : D \cong E : g$ whenever $g \circ f = \mathrm{id}_D$ and $f \circ g = \mathrm{id}_E$. Moreover, we say that D and E are *isomorphic*, and write $D \cong E$, whenever $f : D \cong E : g$ for some f and g.

Prove that for domains D and E, and for monotone functions $f : D \to E$ and $g : E \to D$, if $f : D \cong E : g$ then f and g are continuous. [4 marks]

(*b*) A *refsym* is defined to be a pair $\underline{A} = (A, \sim_A)$ consisting of a set A together with a binary relation on it $\sim_A \subseteq A \times A$ that is reflexive (namely, $x \sim_A x$ for all $x \in A$) and symmetric (namely, $x \sim_A y$ implies $y \sim_A x$ for all $x, y \in A$).

For a refsym $\underline{A} = (A, \sim_A)$, define $\Delta(\underline{A}) = \{\ \alpha \subseteq A \mid \forall\, x, y \in \alpha.\ x \sim_A y\ \}$.

(*i*) Prove that for a refsym $\underline{A}$, the pair $\mathrm{D}(\underline{A}) = (\Delta(\underline{A}), \subseteq)$ is a domain. [5 marks]

(*ii*) For a set A, define a refsym $F(A)$ such that the domain $\mathrm{D}(F(A))$ and the flat domain $A_\perp$ are isomorphic. Establish the isomorphism $\mathrm{D}(F(A)) \cong A_\perp$. [5 marks]

(*iii*) For refsyms $\underline{A}_1$ and $\underline{A}_2$ define a refsym $P(\underline{A}_1, \underline{A}_2)$ such that the domain $\mathrm{D}(P(\underline{A}_1, \underline{A}_2))$ and the product domain $\mathrm{D}(\underline{A}_1) \times \mathrm{D}(\underline{A}_2)$ are isomorphic. Establish the isomorphism $\mathrm{D}(P(\underline{A}_1, \underline{A}_2)) \cong \mathrm{D}(\underline{A}_1) \times \mathrm{D}(\underline{A}_2)$. [6 marks]

5 E-Commerce

Despite the early hope that the internet would help create greater competition and fairer markets, in reality it has given rise to dominant firms in most of the major online markets and hence less competition.

(*a*) Using examples describe five characteristics of online markets that make this statement true. [5 marks]

(*b*) Using examples describe five characteristics of online markets that make this statement false. [5 marks]

(*c*) Discuss if you think that using blockchain based non-fungible tokens will help create fairer online markets, giving reasons for and against? [10 marks]

6 Hoare Logic and Model Checking

Consider a programming language with commands C consisting of the `skip` no-op command, sequential composition C_1`;`C_2, loops `while` B `do` C for boolean expressions B, conditionals `if` B `then` C_1 `else` C_2, assigment X `:=` E for program variables X and arithmetic expressions E, heap allocation X `:= alloc(`E_1`,`$\ldots$`,`E_n`)`, heap assignment `[`E_1`] :=` E_2, heap dereference X `:= [`E`]`, and heap location disposal `dispose(`E`)`. Assume `null` $= 0$, and predicates for lists and partial lists:

$$\begin{aligned}
&\mathrm{list}(t, []) = (t = \mathrm{null}) \wedge emp\\
&\mathrm{list}(t, h :: \alpha) = \exists y.(t \mapsto h) * ((t+1) \mapsto y) * \mathrm{list}(y, \alpha)\\
&\mathrm{plist}(t_1, [], t_2) = (t_1 = t_2) \wedge emp\\
&\mathrm{plist}(t_1, h :: \alpha, t_2) = \exists y.\ (t_1 \mapsto h) * ((t_1+1) \mapsto y) * \mathrm{plist}(y, \alpha, t_2)
\end{aligned}$$

In the following, all triples are linear separation logic triples.

(*a*) Find a command C satisfying the following separation logic partial correctness triple: $\{\top\}\ C\ \{X \mapsto 0 * X \mapsto 0\}$. [2 marks]

(*b*) Give a loop invariant that would serve to prove the following triple, where 'map negate α' is the list of negated values in α (no proof outline required):

$\{\mathrm{list}(X, \alpha)\}$

```
Y = X; while Y ≠ null do (V := [Y]; [Y] = V * (-1); Y = [Y + 1])
```

$\{\mathrm{list}(X, \mathrm{map\ \ negate}\ \ \alpha)\}$ [4 marks]

(*c*) Give a loop invariant that would serve to prove the following triple, for a program that finds the last element of a list (no proof outline required):

$\{\mathrm{list}(X, \alpha\ \texttt{++}\ [l])\}$

```
CUR = X; NEXT = [X + 1];
while NEXT ≠ null do (CUR = NEXT; NEXT = [NEXT + 1]);
LAST = [CUR]
```

$\{\mathrm{list}(X, \alpha\ \texttt{++}\ [l]) \wedge LAST = l\}$ [5 marks]

(*d*) Explain why a proof of Part (*c*) would not succeed if the post-condition of the triple was replaced with $\{emp \wedge LAST = l\}$. [3 marks]

(*e*) Give a loop invariant that would serve to prove the following triple, for a program that copies a given list (no proof outline required):

$\{\mathrm{list}(X, \alpha) \wedge \alpha \neq []\}$

```
V = [X]; Y := alloc(V, null); CUR := [X+1]; OLD = Y;
while CUR ≠ null do (
   V = [CUR]; N = alloc(V, null); [OLD + 1] = N;
   CUR = [CUR + 1]; OLD = N
)
```

$\{\mathrm{list}(X, \alpha) * \mathrm{list}(Y, \alpha)\}$ [6 marks]

7 Information Theory

(*a*) Show how to use Huffman coding to produce optimal *ternary* codewords for a symbol alphabet of size 9 with a uniform probability distribution across the input symbols. Explain how you know it is optimal. [3 marks]

(*b*) A *suffix* code occurs when no codeword is a suffix of any other codeword. For example, `01` precludes `101`. Show that an optimal suffix code exists for every probability distribution over the input symbols. [3 marks]

(*c*) An alternative code assigns a codeword of exact length $\lceil \log_2(\frac{1}{P_i}) \rceil$ to symbol i, which occurs with probability P_i.

(*i*) Explain the significance of $\lceil \log_2(\frac{1}{P_i}) \rceil$ and the logic behind its use in this way. [2 marks]

(*ii*) Can this scheme always produce a prefix code? Justify your answer. [2 marks]

(*iii*) Compare this scheme to a Huffman code. [5 marks]

(*d*) If *all* symbols input to a Huffman code occur with probability $< p$ there can be no codeword of length 1. Find the upper bound for p. [5 marks]

8 Machine Learning and Bayesian Inference

You are playing a game against an opponent who has a biased die with probabilities $\{d_1, \dots, d_6\}$. For each outcome of the die there is a biased coin. The coins show a head with probabilities p_i for $i = 1, \dots, 6$. Your opponent produces a sequence $\mathbf{o} = (o_1, \dots, o_m)$ of heads (H) and tails (T) having length m. Each is generated by first rolling the die, then flipping the coin corresponding to the die's outcome. Let the random variable (RV) denoting the ith outcome be $O_i \in \{H, T\}$, and the RV denoting the outcome of the ith roll of the die be $D_i \in \{1, \dots, 6\}$.

(*a*) Write down an expression for $p_{i,j} = \Pr(O_i | D_i = j)$. [2 marks]

(*b*) Collecting the parameters describing the die and the coins into a vector $\boldsymbol{\theta}$, show that the log-likelihood for the observed outcomes is

$$\log \Pr(\mathbf{o}|\boldsymbol{\theta}) = \sum_{i=1}^{m} \log \sum_{j} p_{i,j} d_j.$$

[4 marks]

(*c*) Define the *latent variables* $z_i^{(j)}$ taking value 1 if the ith outcome is generated by the jth coin and 0 otherwise. Show that

$$\log \Pr(\mathbf{o}, \mathbf{Z}|\boldsymbol{\theta}) = \sum_{i} \sum_{j} z_i^{(j)} (\log p_{i,j} + \log d_j)$$

where $\mathbf{Z}$ collects together all the values for the latent variables. [4 marks]

(*d*) The *Expectation Maximization (EM) Algorithm* defines the expression

$$L(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{\Pr(\mathbf{o}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}$$

for an arbitrary distribution q, and relies on the fact that

$$L(q, \boldsymbol{\theta}) = \log \Pr(\mathbf{o}|\boldsymbol{\theta}) - D_{\text{KL}}(q(\mathbf{Z}) || \Pr(\mathbf{Z}|\mathbf{o}, \boldsymbol{\theta}))$$

where $D_{\text{KL}}(.||.)$ denotes Kullback-Liebler distance. Explain how these expressions lead to the two steps used by the EM algorithm to maximize the likelihood $\log \Pr(\mathbf{o}|\boldsymbol{\theta})$. [5 marks]

(*e*) Derive the *E step* of the EM algorithm for maximizing the likelihood in the case of the die and coins problem. Your answer should include an expression for the resulting probability distribution in terms of the parameters $\boldsymbol{\theta}$. [5 marks]

9 Optimising Compilers

A language $\mathcal{L}$ has the following abstract syntax, where c ranges over integer constants, x ranges over a set of variables and $\oplus$ ranges over binary operations:

$$e = c \mid x \mid \lambda x.e \mid e_1 e_2 \mid \texttt{let } x = e_1 \texttt{ in } e_2 \mid \texttt{if } e_1 \texttt{ then } e_2 \texttt{ else } e_3 \mid e_1 \oplus e_2$$

Consider the following program P in $\mathcal{L}$:

```
let x = 5 in
    let f = λx.2 ∗ x in
        if x > 0 then f x else f (0 − x)
```

This question asks you to perform 0CFA on P.

(*a*) Draw the program P as a tree and label its program points. [4 marks]

(*b*) Give the space of flow values for P. [2 marks]

(*c*) Each program point i in P has an associated flow variable α_i. Show the initial constraints on each α_i that are generated when performing 0CFA. [4 marks]

(*d*) Show how the process of solving the constraints from part (*c*) leads to additional constraints being generated. [4 marks]

(*e*) Show the final solution after solving all constraints from parts (*c*) and (*d*) and simplifying binary terms. [4 marks]

(*f*) Explain whether your answer is a safe over- or under-approximation of the result of P and where the imprecision comes from. [2 marks]

10 Principles of Communications

(*a*) Imagine you are set the task to optimise road traffic, so that journey times are minimised for a given set of vehicles moving between a known set of sources and destinations. You have the freedom to fix routes and to control speed limits on routes. How can you tackle this problem? [5 marks]

(*b*) Now imagine we have roads with electronic signage that allows us to declare variable speed limits. What approach could you take to setting the speeds to minimise congestion? [5 marks]

(*c*) In an effort to reduce traffic congestion, the government decides to introduce a charge for using the roads during the busiest times of day. Explain how this may lead to an increase in the overall welfare of all drivers. Answers should include discussion of concepts such as willingness to pay, and peak rate charging. [5 marks]

(*d*) Without imposing strict routes, each car driver chooses a path independently, but what about a delivery fleet of vans/trucks? Perhaps parts of the road network are set aside for a known set of flows of haulage vehicles. How could this differentiation between individual and groups of vehicles be deployed? [5 marks]

11 Quantum Computing

(*a*) A Toffoli gate is to be used as the oracle in the Deutsch-Jozsa algorithm.

(*i*) Why is this not a valid oracle for the Deutsch-Jozsa algorithm? [1 mark]

(*ii*) If the Deutsch-Jozsa algorithm is run anyway with a Toffoli gate as the oracle, what will the outcome be? [6 marks]

(*iii*) How can two Toffoli gates be used to construct an oracle that *is* valid for the Deutsch-Jozsa algorithm? [2 marks]

(*b*) Give a (single qubit) quantum circuit that can perfectly distinguish the states $|+\rangle$ and $|-\rangle$ using any unitary operations, but only computational basis measurements. [2 marks]

(*c*) Show that the quantum states

$$\frac{1}{\sqrt{2}}\left(|+\rangle + |-\rangle\right) \quad \text{and} \quad \frac{1}{\sqrt{2}}\left(|+\rangle - |-\rangle\right)$$

can be perfectly distinguished. Give the measurement basis to achieve this in terms of the computational basis states $|0\rangle$ and $|1\rangle$. [3 marks]

(*d*) Let $|\psi\rangle$ be some unknown quantum state, which is either $|1\rangle$ or $\frac{\sqrt{3}}{2}|0\rangle + \frac{1}{2}|1\rangle$. Furthermore it is known that there is a 75% probability that $|\psi\rangle$ is $|1\rangle$ and a 25% probability that $|\psi\rangle$ is $\frac{\sqrt{3}}{2}|0\rangle + \frac{1}{2}|1\rangle$.

A measurement must be performed to help identify which state $|\psi\rangle$ is. Give a measurement basis that guarantees to correctly determine $|\psi\rangle$ for one of the measurement outcomes; if there are multiple such bases, give the one that maximises the overall probability of correctly identifying $|\psi\rangle$. Give the probability of success. [6 marks]

12 Randomised Algorithms

(*a*) Consider the following Markov chain with state space $\Omega = \{1, 2\}$ and transition matrix:

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix},$$

where $p \in [0, 1]$ and $q \in [0, 1]$.

(*i*) For the class of Markov chain above, state whether an instance: (1) is irreducible, (2) is aperiodic and (3) has a unique stationary distribution. Pay attention to special cases. [8 marks]

(*b*) Consider now the transition matrix:

$$P = \begin{pmatrix} 5/6 & 1/6 \\ 1/3 & 2/3 \end{pmatrix}.$$

(*i*) Prove that for any integer $k \geq 1$, the k-th power of P satisfies:

$$P^k = \begin{pmatrix} 2/3 & 1/3 \\ 2/3 & 1/3 \end{pmatrix} + (1/2)^{k-1} \cdot \begin{pmatrix} 1/6 & -1/6 \\ -1/3 & 1/3 \end{pmatrix}$$

[4 marks]

(*ii*) State the general definition of the mixing time $\tau(\epsilon)$ of a Markov chain with transition matrix P. [2 marks]

(*iii*) Consider now again the transition matrix P from (*b*). What can you deduce for $\tau(1/24)$?
Hint: You may use the formula from (*b*)(*i*). [6 marks]

13 Types

(*a*) Consider the OCaml option type

```
type 'a option = None | Some of 'a
```

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

(*i*) For a fixed A, give a suitable System F type for a Church encoding of the `A option` type. [1 mark]

(*ii*) Give an implementation of the `Some` and `None` constructors for this encoding. [2 marks]

(*iii*) Give a type and encoding of an eliminator named `case` for the option type. [2 marks]

(*iv*) Give the reduction rules for `case`, and show that your encoding models them correctly. [5 marks]

(*b*) All of the questions in this part are about the monadic lambda calculus.

(*i*) Give a well-typed term of type $T(T(A)) \to T(A)$, and explain briefly in prose what this function does. [2 marks]

(*ii*) Give a well-typed term of type $T(A) \to (A \to T(B)) \to T(B)$, and explain briefly in prose what this function does. [2 marks]

(*iii*) Give a type and definition of a monadically-typed fixed point operator suitable for defining recursive functions on integers. [6 marks]

END OF PAPER