IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2017

MEng Honours Degrees in Computing Part IV
MSc in Advanced Computing
MSc in Computing Science (Specialist)
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant examinations for the Associateship of the City and Guilds of London Institute

PAPER C436H

PERFORMANCE ENGINEERING

Wednesday 22 March 2017, 11:40
Duration: 70 minutes

Answer TWO questions

Paper contains 3 questions
Calculators required

1 a Consider the following probability transition matrix

$$P = [p_{ij}] = \begin{array}{c} \\ E \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ X \end{array} \begin{array}{c} \begin{array}{cccccc} E & S_1 & S_2 & S_3 & S_4 & X \end{array} \\ \begin{bmatrix} 0 & 1.0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.4 & 0.2 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.0 \\ 0 & 0.5 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 1.0 \end{bmatrix} \end{array}$$

describing user interactions with an IT system hosting services S_1, S_2, S_3, S_4, and where E and X respectively denote entry and exit states.

i) Determine the mean session length.

ii) Give a formula to compute the probability of ending the session after exactly three service calls. (You are *not* asked to compute this probability.)

iii) Assume that the application has $N = 10$ users, each starting a new session at a rate of $\lambda = 0.11$ sessions/sec. Services are hosted on $M = 2$ virtual machines (VMs) and require the following service times:

Time [s]	VM 1	VM 2
S_1	1.0	0.2
S_2	1.0	0.3
S_3	1.0	0.3
S_4	1.0	0.3

If the load balancer uses a round-robin dispatching policy, what is the expected CPU utilization at each VM?

b Discuss the differences between batch and interactive workloads. Then present the multi-tier architecture model for J2EE applications.

The two parts carry, respectively, 70% and 30% of the marks.

2 a Consider the following four configuration options for the Java Virtual Machine:

- A: number of garbage collector threads, with levels 4 and 8 threads.
- B: garbage collector heap size, with levels 16Mb and 32Mb.
- C: serial garbage collection, with levels ON and OFF.
- D: pause time between collections, with levels 50ms, 100ms, 150ms.

i) Propose a 2^{4-1} design to determine an optimal configuration with respect to response time. Write down *any two* rows of the sign table, deriving all the confoundings and the design resolution.

ii) Suppose that a 2^2 design is carried out for factors B and C, leading to the following measurements:

Response time (s)	OFF	ON
16Mb	11	9
32Mb	8	16

Quantify the percentages of variation explained by each factor and by their interaction.

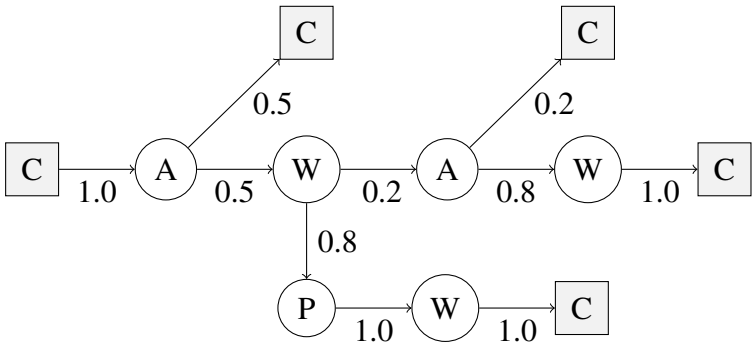
iii) Explain the interpretation of the effects q_0, q_B, q_C, and q_{BC}.

iv) Write down the response models underpinning the 2^2 and 2^{4-1} designs.

b Using SPECjbb2015 and Apache JMeter as reference examples, present the main similarities and differences between industry benchmarks and load testing tools.

The two parts carry, respectively, 70% and 30% of the marks.

3 a Consider the following client-server interaction diagram for a client (C), a web server (W), an application server (A), and a proxy server (P):



i) List the possible transaction flows of the client and determine their probability.

ii) Each call to W, A, and P requires an average service time of $10ms$, $30ms$ and $50ms$, respectively. Let workload class 1 consist of transactions that do not call P, while class 2 consists of the rest. Determine the demands at W, A, and P for the two classes.

iii) Find the potential bottlenecks for this system. Can the system incur bottleneck switches? Justify your answer.

iv) Now suppose that class 2 requests arrive at rate $\lambda_2 = 10tps$, give a bound on the sustainable arrival rate of class 1 requests.

b i) Explain the main differences between the IaaS, PaaS, and SaaS cloud service models.

ii) Give an example of autoscaling rule based on static thresholds.

The two parts carry, respectively, 70% and 30% of the marks.