IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2019-2020

MSc in Computing Science (Specialist)
MSc in Advanced Computing
MEng Honours Degrees in Computing Part III
MEng Honours Degree in Mathematics and Computer Science Part III
BEng Honours Degree in Mathematics and Computer Science Part III
MEng Honours Degree in Electronic and Information Engineering Part IV
BEng Honours Degree in Electronic and Information Engineering Part III
MEng Honours Degree in Electronic and Information Engineering Part III
BEng Honours Degree in Computing Part III
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 C337

SIMULATION AND MODELLING

Wednesday 11th December 2019, 10:00
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators required

1 A high-street post office has a fixed population of C customers who visit the office on average every Δt time units. The post office has two service counters and a waiting area that can support up to N waiting customers who are served in First-Come First-Served order. If an arriving customer finds the waiting room full they leave immediately, returning Δt time units later on average. Otherwise the customer is served at one of the two counters, possibly after waiting in the queue. Customers in service are not considered to be waiting.

One counter is permanently open, but the second is only opened when there are M or more waiting customers, for some parameter $1 \leq M \leq N$. When a customer at the second counter completes their service, the counter closes if the total number of waiting customers at that time is less than M.

a Design a discrete-event simulation for this system by specifying the state variables and events required. Sketch the code for each event using any notation you see fit. State any assumptions you make.

b Outline how you would augment the simulation with measurement code for estimating the probability, $p(k)$ say, that the number of waiting customers is k, for a given $0 \leq k \leq N$.

c Now suppose that we use the simulation to compute an unbiased estimate of $p(k)$ at equilibrium and that the measures you identified for part b are reset after a suitably chosen warm-up period, w. Given that the state at time w might be precisely that at time 0, how does the warm-up help to ensure that the estimate is unbiased?

d With $C = 10000$ and Δt = 7 days (= 10080 minutes), the system was measured for 8hrs (= 480 minutes). The mean service time for both counters was 1.25 minutes. There were 327 completions at the first counter, 148 at the second and the average number of waiting customers was 1.52. All customers who entered the post office eventually left. What was the average response time and the average total population of the post office, including both queueing customers and those being served? Show your working.

The four parts carry, respectively, 50%, 15%, 15%, and 20% of the marks.

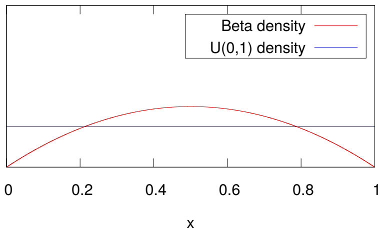
2 a In what sense is the exponential distribution “memoryless”? As part of your answer show that if X is an exponentially-distributed random variable then $P(X \leq t + s \mid X > t) = P(X \leq s)$, for $t > 0, s > t$.

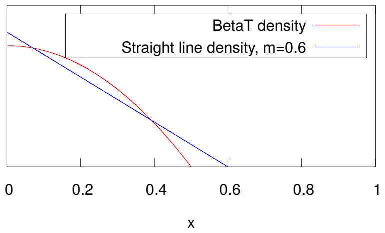
b A discrete-event simulation works by scheduling future events onto a time-ordered priority queue. If all new events are scheduled at some exponentially-distributed time in the future, how might it be possible, in principle, to obviate the need for the priority queue?

c A particular Beta distribution has density function and cumulative distribution function (cdf) respectively given by:

$$f(x) = 6x(1-x) \quad F(x) = 3x^2 - 2x^3 \quad (0 \leq x \leq 1)$$

Explain how the acceptance-rejection (AR) method can be used to sample the distribution using a scaled uniform density function on the interval $(0, 1)$ as the dominating function (see the left figure below).





d In an attempt to improve the efficiency of the AR sampler from part c, it is proposed to work with a truncated version of the distribution, “BetaT”, say, where the right half of the Beta density is mapped to the interval $(0, \frac{1}{2})$. The idea is to use the scaled density function of a “straight line” distribution as the dominating function (see the right figure above) and to sample that distribution using the inverse transform method. The density function and cdf for the straight line distribution to be used are respectively $g(x) = \frac{2}{m^2}(m - x)$ and $G(x) = \frac{2x}{m} - \frac{x^2}{m^2}$, with $m = 0.6$.

i) Show that the density function of the BetaT distribution is $f(x) = 3 - 12x^2, \; 0 \leq x \leq 0.5$.

ii) How would you transform a sample from the BetaT distribution into a sample from the Beta distribution?

iii) Compute the mean number of $U(0, 1)$ random samples required by the AR method in both cases. Which of the two methods do you think will be the most efficient overall? Justify your answer.

The four parts carry, respectively, 20%, 15%, 20%, and 45% of the marks.

3 a An M/M/1/2 queue, i.e. an M/M/1 queue with capacity 2, including the job in service, has service rate μ and an arrival rate of either λ_1 or λ_2. The arrival rate changes only when the server is full, whereupon it switches as a Markov process with instantaneous transition rate α from λ_1 to λ_2 and β from λ_2 to λ_1.

i) Define the state space of this queue, draw its state-transition diagram and indicate its instantaneous transition rates.

ii) Hence write down its steady state probabilities, up to a normalising constant which you need not calculate.

iii) Show that the queue is equivalent to a standard, state-dependent M/M/1/5 queue and specify its service and arrival rate functions.

b i) Show that if you merge two Poisson streams with rates λ_1 and λ_2 the merged process is Poisson with rate $\lambda_1 + \lambda_2$

ii) Illustrate, with an example, the splitting properties of the Poisson process. How does the output of a fork node differ from a probabilistic split if the arrival process to the fork is Poisson?

iii) Define the nth harmonic number H_n. Then obtain an expression for the expected value of the maximum of the n identical exponential random variables in terms of H_n.

The two parts carry, respectively, 55%, and 45% of the marks.

4 In a closed queueing network of M nodes, with a serial topology, and having population K, node m has constant service rate μ_m and relative load x_m $(1 \leq m \leq M)$.

a i) Write down a general expression for the balance equations of this model. In your answer, define n_m to be the number of jobs in node m and δ_m to be 1 if $n_m > 0$ and 0 otherwise, $(1 \leq m \leq M)$.

ii) Define the normalising constant function $g(k, m)$, $1 \leq k \leq K$, $1 \leq m \leq M$, for this network's equilibrium state probabilities. What is the relationship between μ_m and x_m?

b Assume in this part that $M = 2$, $K = 3$, $x_1 = 5$, $x_2 = 1$.

i) Using a suitable recurrence relation for the normalising constant, compute the value of $g(K, M)$.

c Assume in this part that $x_1 = x_2 = \ldots = x_M$.

i) Express the normalizing constant as a function of K, M and x_1 only.

ii) Obtain an explicit formula for the network throughput $T(K, M)$. Using this, determine the utilization of the nodes when K and M tend to infinity while keeping the ratio $c = M/K$ constant.

The three parts carry, respectively, 35%, 25%, and 40% of the marks.

final

IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2017-2018

BEng Honours Degree in Computing Part III
BEng Honours Degree in Electronic and Information Engineering Part III
MEng Honours Degree in Electronic and Information Engineering Part III
BEng Honours Degree in Mathematics and Computer Science Part III
MEng Honours Degree in Mathematics and Computer Science Part III
MEng Honours Degrees in Computing Part III
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 C337

SIMULATION AND MODELLING

Monday 11 December 2017, 14:00
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators not required

1 a Assume that a system is observed for T seconds during which time there were n job completions. Let the population at time $0 \leq t \leq T$ be $N(t)$ jobs and $R_i, 1 \leq i \leq n$, be the response time of the i^{th} job.

i) If $N(0) = N(T) = 0$ explain, with the aid of a diagram, why the area under $N(t)$, i.e. $\int_0^T N(t)\, dt$, is equal to $\sum_{i=1}^{n} R_i$. Assume a particular queueing discipline, e.g. first-come-first-served.

ii) It is possible to show that $\frac{1}{T} \int_0^T N(t)\, dt$ converges to the mean population, N, as $T \to \infty$. Using this, and the property from part i, derive Little's law, viz. $N = XR$, where X is the throughput and R the average response time.

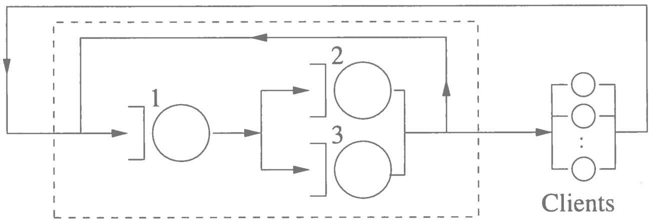


Fig. 1: A client-server system

b Consider the (closed) system in Figure 1, which comprises N clients who submit requests to a three-node server, indicated by the dashed box. On average each client request makes 300 visits to node 1, 180 to node 2 and 120 to node 3. The corresponding average service times are 10ms, 20ms and 30ms respectively and the average client think time is $Z = 30s$.

i) Determine the bottleneck device, or devices, and compute the maximum throughput that can be achieved by the server for both low load ($N = 1$) and high load ($N \to \infty$). Show your working.

ii) Now suppose that the server is optimised so that the average service time of node 3 is reduced by 5%. Under high load what is the maximum throughput that can now be achieved and what will happen to the average queue populations as $N \to \infty$? Explain your answers.

iii) Outline a discrete-event simulation of the system. As part of your answer identify the events required and explain how you would model the system state and the branching in the network. You should make it clear what each event does, although you are not required to spell out the event code in detail.

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

2a A particular *triangular* distribution on the interval $0 \le x \le b$ has a density function given by:

$$f(x) = \begin{cases} \frac{2x}{bc}, & 0 \le x \le c \\ \frac{2(b-x)}{b(b-c)}, & c \le x \le b \end{cases}$$

i) Show how the inverse transform method can be used to sample this distribution. *Hint*: If random variable U is uniformly distributed on the interval $(0,1)$ then consider the two sub-intervals $0 < U \le F(c)$ and $F(c) < U < 1$ separately, where $F(x)$ is the cumulative distribution function.

ii) With the aid of a diagram show how the rejection method can be used to sample the distribution using a scaled uniform distribution on the interval $(0, b)$ as the dominating function. How many random ($U(0,1)$) numbers will it take on average to sample the distribution using this method?

b Consider a Markov Process with N states and generator matrix $\mathbf{Q}$ with off-diagonal elements $q_{i,j}, 1 \le i, j, \le N, i \ne j$ and diagonal elements $q_{i,i} = -\sum_{1 \le j \le N, j \ne i} q_{i,j}, 1 \le i \le N$. Also, let $\mathbf{p} = (p_1, p_2, ..., p_n)$ be the vector of equilibrium (steady state) probabilities, where p_i is the probability that the system is in state $i, 1 \le i \le N$.

i) Show that the holding time of state $1 \le s \le N$ is exponentially distributed with parameter $-q_{s,s}$.

ii) Write down the general form for the global balance equations and explain why they are equivalent to the system of equations $\mathbf{pQ} = \mathbf{0}$.

iii) What is meant by the *memoryless* property of the exponential distribution? As part of your answer explain why, when simulating a Markov Process, it is not necessary to use an explicit event list (priority queue) to store state transition events which have yet to be invoked.

The two parts carry equal marks.

3 a Consider a set of n items having identical sizes and a cache with a capacity of $m \leq n$ items, $m \geq 1$. Requests for item k arrive according to a Poisson process with rate λ_k, $1 \leq k \leq n$. A *cache hit* occurs if the item is stored in the cache; otherwise the request results in a *cache miss*. The cache uses a least-recently used (LRU) replacement policy, which behaves as follows:

- *Cache hit*: If a request for item k finds it in cache position $p > 1$, then item k is moved into position 1 and all the items in positions $1, \dots, p-1$ are shifted back by one position, i.e., into positions $2, \dots, p$ and item k is moved into position 1. If the request finds the item k in position $p = 1$, no changes are made.

- *Cache miss*: If a request for item k does not find it in the cache, then the item in position m is removed from the cache. Items in position $p = 1, \dots, m-1$ are thenshifted into positions $p' = 2, \dots, m$, and item k is added to the cache in position 1.

Assume that $m = 2$ and $n = 3$.

i) Suppose one were to model the cache state by a CTMC with state $n = (n_1, \dots, n_m)$, where $n_i = k$ if and only if item k occupies position i in the cache and $n_i \neq n_j$ if $i \neq j$. Write the global balance equations for this CTMC.

ii) Using the result in part i), prove that the cache miss probability is equivalently given by the expression

$$p_{miss} = \lambda_1\lambda_2\lambda_3 \left(\frac{1}{\lambda_1 + \lambda_2} + \frac{1}{\lambda_1 + \lambda_3} + \frac{1}{\lambda_2 + \lambda_3} \right)$$

iii) Using the steady-state probabilities of the CTMC, compute the miss rate of the cache when $\lambda_k = 2^{-k}$. Compare the result with the expression of p_{miss} given in part ii).

b State the PASTA property. Then use it show that the response time distribution of a M/M/1 queue with arrival rate λ and service rate μ is $f_R(x) = (\mu - \lambda)e^{-(\mu-\lambda)x}$. (*Hint*: the density of an Erlang-$(j+1)$ distribution is $e_{j+1} = \mu\frac{(\mu x)^j}{j!}e^{-\mu x}$).

The two parts carry, respectively, 75%, and 25% of the marks.

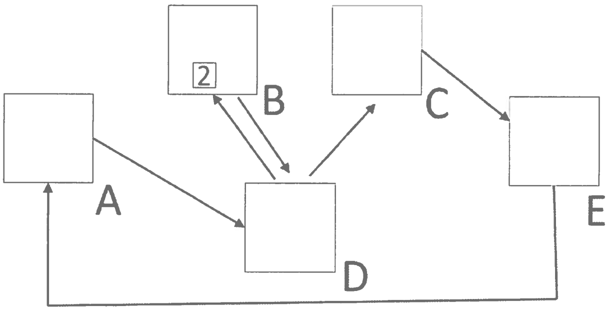
4a A network engineer is studying $M = 2$ network links arranged in tandem. Each link can be modelled as a M/M/1 first-come first-served queue. Packets arrive according to a Poisson process with rate γ and are served with identical exponential rate μ.

i) Determine the mean queue-length N and the mean waiting buffer length N_Q at each node when $\mu = 1$ and $\gamma = 0.6$.

ii) Determine a formula for the probability that the *total* number of jobs inside the network is exactly m jobs. How would the formula change if the queues have different service rates? (*Hint*: $\sum_{n=m}^{\infty} a^n = a^m(1-a)^{-1}$ if $a < 1$.)

iii) Suppose now that each link can process multiple packets in parallel so that each node may be seen as a queue with infinite servers. Knowing that the departure process of a $M/M/\infty$ queue is a Poisson process, determine an expression for the joint state probabilities of the network.

b Consider the following logical processes (LPs) running in a parallel simulation *without* null messages:



Assume that all LPs are at virtual time 0. Each event processed by a LP generates a new event on all outgoing links. The initial buffer contents are $AD = \{\emptyset\}$, $BD = \{\emptyset\}$, $CE = \{\emptyset\}$, $DB = \{2\}$, $DC = \{\emptyset\}$, and $EA = \{\emptyset\}$.

i) Using the deadlock-prone conservative simulation algorithm, determine the simulation sequence for the model in the figure above when the distance between any pair of connected LPs is 1. Terminate the sequence after the first round.

ii) Show how the simulation sequence changes using the null message algorithm. Use a lookahead equal to $L = 1$ and terminate the sequence at the end of the second round.

The two parts carry equal marks.

IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2017

BEng Honours Degree in Computing Part III
BEng Honours Degree in Electronic and Information Engineering Part III
MEng Honours Degree in Electronic and Information Engineering Part III
BEng Honours Degree in Mathematics and Computer Science Part III
MEng Honours Degree in Mathematics and Computer Science Part III
MEng Honours Degrees in Computing Part III
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 C337

SIMULATION AND MODELLING

Monday 12 December 2016, 14:00
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators required

1 a An interactive system has N users who submit requests to a server comprising a preprocessing unit (P) and a compute engine (E). The system has been monitored for one hour (T=3600 seconds) during which time there were 1200 completions in total. The average user think time was $Z = 10$ seconds and the busy periods and completion counts for the two devices, P and E, were respectively $B_P = 2200$, $C_P = 1200$, $B_E = 3400$, $C_E = 120000$.

i) Compute the visit counts (V), mean service times (S), service demands (D) and utilisations (U) for the two devices.

ii) Which is the bottleneck device and what is the upper bound on throughput that can be achieved under heavy load? Show your working.

iii) Compute the lower bound on the average waiting time that can be experienced under light load.

iv) Now suppose that you try to fix the bottleneck by purchasing a duplicate copy of the bottleneck device and arranging for its original demand to be spread evenly across the two duplicates. What is the upper bound on throughput and the lower bound on average waiting time that can now be achieved? Show your working.

b The Weibull distribution has density and cumulative distribution function (cdf):

$$\begin{aligned} f_W(x) &= \frac{\beta x^{\beta-1} e^{-(x/\alpha)^\beta}}{\alpha^\beta} \qquad \alpha, \beta > 0 \text{ and } x \geq 0 \\ F_W(x) &= 1 - e^{-(x/\alpha)^\beta} \end{aligned}$$

i) Show how the inverse transform method can be used to sample from a Weibull distribution. As part of your answer prove the correctness of the method by showing that if $U \sim U(0,1)$ then $X = F_W^{-1}(U)$ has cdf F_W.

ii) Assume you are able to sample a Weibull distribution, e.g. using the method from part i). Explain how the *acceptance-rejection* method can be used to sample a Gamma distribution using the Weibull distribution with parameters $\alpha = 2, \beta = 1$ as the dominating function. The Gamma distribution to be sampled has density function given by:

$$f_G(x) = 4xe^{-2x}, \quad x \geq 0$$

As part of your answer determine a scaling factor, c, that ensures that $cf_W(x) \geq f_G(x)$ for all $x \geq 0$.

The two parts carry, respectively, 45% and 55% of the marks.

2 Consider a multi-core shared-memory system with N cores, each executing one thread that is pinned to the core. The threads alternate between executing critical and non-critical sections using a ticket spinlock, as illustrated by the following thread code:

```
repeat indefinitely
  <Non-critical work section...>
  // Claim lock...
  myTicketId = atomicFetchAndInc(GlobalTicketId);
  while(currentOwnerTicketId != myTicketId) {};
  <Critical section...>;
  // Release lock...
  atomicIncrement(currentOwnerTicketId);
```

The average times spent executing the critical and non-critical sections are t_c and t_x respectively. When a thread releases the lock each cached copy of `currentOwnerTicketId`, will be updated by the hardware one at a time and each update takes time t_u on average.

In the coursework exercise we assumed that the releasing thread resumes its non-critical section at the same time that the next thread acquires the lock, i.e. *after* the update to `currentOwnerTicketId` has completed. Now assume that the releasing thread resumes its non-critical section *immediately* after executing its critical section (average time t_c), i.e. without waiting for the updates to cached copies of `currentOwnerTicketId` to complete.

a Assuming that all times delays are exponentially distributed, draw a Continuous Time Markov Chain (CTMC) for modelling this system for the case $N = 3$ and explain its structure. Note that it is now possible for the releasing thread to attempt to re-acquire the lock before it has been re-assigned to the next waiting thread (if there is one); your model should reflect this. Specify clearly the transition rates in terms of t_x, t_c and t_u and justify any assumptions you make.

b For your CTMC in part (a) determine the generator matrix (**Q**) and show how it can be transformed so as to have rank equal to the number of states in the CTMC.

c Now suppose that the various time delays are not exponentially distributed. Design a discrete-event simulation of the system that estimates the average number of critical sections executed per unit time for a given $N > 0$. You may assume the existence of methods for sampling the various time delay distributions and, in particular, a method `sampleU(n)` for sampling the time taken to update `currentOwnerTicketId` when there are `n` threads queued waiting to acquire the lock (i.e. you do not need to model the individual messages required to implement the update). Specify how you would initialise the simulation and state any assumptions you make in designing your model.

The three parts carry, respectively, 40%, 15%, and 45% of the marks.

3 a Consider a web server modelled as an infinite-capacity M/M/1 queue with arrival rate λ and service rate μ. Assume that, for maintenance, the operating system periodically suspends the web server from execution, resuming it after a short amount of time. When suspended, the web server state cannot be modified.

i) Draw the state space of a Markov process that models this system. Call μ' the rate at which the web server is suspended, and λ' the rate at which it is resumed.

ii) Determine the global balance equations (GBEs) for the Markov process obtained in part i). Denote by $p(n)$ the equilibrium probability of being in a state where the web server is active and $n \geq 0$ jobs are enqueued, and by $p(n')$ $(n' \geq 0)$ the corresponding probability when the web server is suspended.

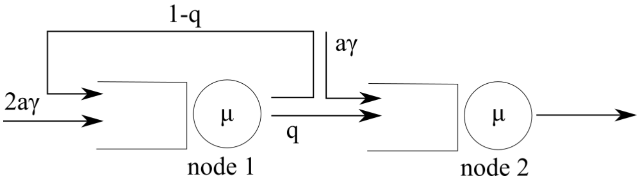
iii) Using the GBEs, express $p(n)$ as a function of $p(n')$. Use the result to reformulate the GBEs in terms of $p(n')$ only.

iv) Verify that the analytical expression of $p(n')$ is given by

$$p(n') = C(1-\rho)^{-1}\rho^{n'} \qquad n' \geq 0$$

where $\rho = \lambda/\mu$ and C is a normalizing constant.

b The following queueing network represents a transaction processing system where jobs arrive at rate $2a\gamma$ at the first node and $a\gamma$ at the second node, where a can be modified to avoid overloading subject to $a > 0$. The system has two nodes, each with service rate μ. The internal branching probabilities between the service centers are shown on the diagram.



i) Find the conditions on a for the network to be stable.

ii) Taking $a = \frac{1}{2}$ and $q = \frac{2}{3}$ and assuming each of the service nodes is an M/M/1 queue, find the mean number of jobs in the network at equilibrium.

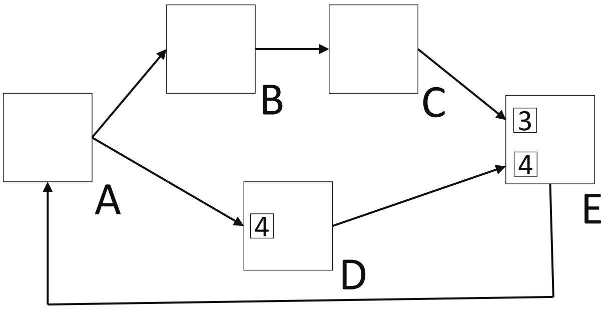
The two parts carry, respectively, 60% and 40% of the marks.

4 a i) State and prove the memoryless property of the exponential distribution.

ii) Illustrate, with an example, the forking and merging properties of the Poisson process.

iii) Consider three Poisson processes each with rate $\lambda = 2$ job/s, feeding a join node. Assume that the join node is initially empty and emits a job only when it receives at least one job from each of the three Poisson processes. What is the expected time before the join node fires the first job?

b Consider the following logical processes (LPs) running in a parallel simulation *without* null messages:



Assume that all LPs are at virtual time 0, except LP A that is at virtual time 1. Assume that A can send event messages only to D. The initial buffer contents are $AB = \{\emptyset\}$, $BC = \{\emptyset\}$, $AD = \{4\}$, $DE = \{4\}$, $CE = \{3\}$, and $EA = \{\emptyset\}$.

i) Give a definition of the local causality constraint (LCC) used in parallel and distributed simulation.

ii) Using the deadlock-prone conservative simulation algorithm, determine the simulation sequence for the model in the figure above when the distance between any pair of connected LPs is 3. Terminate the sequence when it reaches a deadlock.

iii) Show how the simulation sequence changes using the null message algorithm. Null messages generated by A are sent to both B and D. Use a lookahead equal to $L = 3$ and terminate the sequence after the first round.

The two parts carry, respectively, 35% and 65% of the marks.

IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2015-2016

BEng Honours Degree in Computing Part III
BEng Honours Degree in Electronic and Information Engineering Part III
MEng Honours Degree in Electronic and Information Engineering Part III
BEng Honours Degree in Mathematics and Computer Science Part III
MEng Honours Degree in Mathematics and Computer Science Part III
MEng Honours Degrees in Computing Part III
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 C337

SIMULATION AND MODELLING

Friday 18 December 2015, 10:00
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators required

1 A network controller maintains a buffer that the operating system (OS) uses to store packets prior to their transmission across a network. The buffer has the capacity to store up to c packets. Periodically, the controller performs a "flip" operation that gives the OS the illusion that the buffer it has been filling has been instantaneously emptied. After a flip operation the OS continues filling the (now empty) buffer with new packets as before and the controller transmits the previously buffered packets over the network. (In practice this is achieved by *double buffering*, but the details are not important for the exercise.) If the buffer is full when the OS wants to write a packet to it the packet is dropped.

a Outline a discrete-event simulation of this buffering system that will estimate the probability of a packet being dropped and the *distribution* of the number of packets in the buffer, e.g. in the form of an array of probabilities, one for each population $0 \leq i \leq c$. The model should be of the buffer, as perceived by the OS, i.e. a single buffer whose population is either incremented by one by the OS, or set to zero by the controller. Assume that the times between the buffer being emptied are random variables that are independent of the current state of the buffer. Assume also that the time taken to add a packet to the buffer and the time taken to empty (flip) the buffer are negligibly small in comparison to the other time delays in the model and so can be ignored. As part of your answer identify the states and events required, sketch the code for each event and state how the model will be initialised. State any additional assumptions you make. You may use any notation you wish.

b Assuming that the times between packets being added to the buffer and the times between two consecutive emptying (flip) events are both exponentially distributed with respective rates λ and μ, draw a Markov Process (CTMC) that models the state of the buffer and write down the (global) balance equations. Assume that the interface will only attempt to empty (flip) the buffer when there is at least one packet in the buffer.

c Show that the balance equations in part b above have the analytical solution:

$$\begin{aligned}
p_0 &= \frac{\mu}{\lambda+\mu} \\
p_n &= \left(\frac{\mu}{\lambda+\mu}\right)\left(\frac{\lambda}{\lambda+\mu}\right)^n \\
p_c &= \frac{\lambda}{\mu}\left(\frac{\mu}{\lambda+\mu}\right)\left(\frac{\lambda}{\lambda+\mu}\right)^{c-1}
\end{aligned}$$

Hint: Notice that $p_0 = 1 - \sum_{n=1}^{c} p_n$.

The three parts carry, respectively, 40%, 25%, and 35% of the marks.

2a In an interactive system a number of clients submit jobs to a server that comprises a database engine and a processor. 200 user job completions were observed in an observation period of 140 seconds during which time there was a fixed population of 20 clients. There were 1260 completions at the processor and 810 at the database engine and the mean service times at the two devices were 50ms (0.05s) and 130ms (0.13s) respectively. The average client think time was 5 seconds and there is a single class of jobs.

i) Compute the utilizations of the two devices over the observation period and determine the bottleneck device.

ii) What was the average response time of the server during the observation period?

iii) In order to cope with an increasing number of users it has been proposed that the bottleneck device be duplicated so that there are two identical devices sharing the load that is currently being imposed on the original. If both duplicates experience half the average service demand of the original device, what will be the maximum throughput that can be achieved for large N? Show your working.

iv) For the scenario in part iii above, sketch the throughput bounds for both the original configuration and that with the duplicated device (part iii) on the same plot.

b A particular type of *gamma* distribution has a density function given by:

$$\frac{1}{(k-1)!\,\theta^k}\,x^{k-1}e^{-x/\theta} \qquad x \geq 0$$

where $\theta > 0$ and $k > 0$ are parameters and k is an *integer*.

i) Describe the inverse transform method for sampling a distribution and show how the method can be used to sample a gamma distribution, as defined above, for the special case where $k = 1$. As part of your answer prove the correctness of the method.

ii) For the case where $k = \theta = 2$ explain, with the aid of a diagram, how the *acceptance-rejection* (AR) method can be used to sample the distribution by using a scaled exponential density function $h(x) = cg(x) = c\lambda e^{-\lambda x}$, for some $c, \lambda > 0$, as the dominating function. As part of your answer determine a value for c that ensures that $h(x) \geq f(x)$ for all $x \geq 0$.

iii) For the AR method in part ii above determine the average number of $U(0,1)$ random numbers required to produce one sample from the distribution. Explain your answer.

The two parts carry, respectively, 45%, and 55% of the marks.

3a A fraud detection application analyzes tax forms. Each form is processed in a single cycle of computation. Each cycle consists of two sequential operations: (i) a read from a cache at a rate of 1 form/msec; (ii) a processing operation performed by a single server at a service rate of 10 forms/msec. The application can run 6 worker threads in total and each thread will handle a single form at a time.

i) Assume that the response time of a job at the cache does not depend on the number of jobs in the cache. Draw a closed queueing network that models this application.

ii) Use $I = 2$ iterations of the approximate mean-value analysis (AMVA) algorithm to estimate the mean throughput of completed forms.

iii) Suppose now that an algorithm will assign a type to each form according to its content. Types will be "citizen" and "enterprise". Citizen forms and enterprise forms will require 1 and 5 dedicated worker threads, respectively. Enterprise forms will have higher processing priority at the server than citizen forms. The service rates are unchanged and are identical for the two form types. Assuming the service times are all exponentially distributed, use the priority AMVA algorithm with $I = 2$ iterations to determine the mean response time of citizen forms.

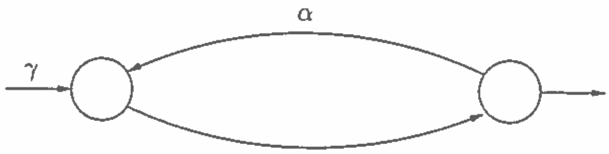
b Consider the theory of fork-join systems:

i) Define the nth harmonic number H_n. Then obtain an expression for the expected value of the maximum of n independent and identically distributed exponential random variables in terms of H_n.

ii) Write the approximate mean-value analysis (AMVA) equations needed to analyze a closed tandem network with K jobs and consisting of a multi-server queue with C servers, each having service demand D_1, and a fork-join subsystem with n parallel stations, each having demand D_2. Explain how you would solve these equations.

The two parts carry, respectively, 65%, and 35% of the marks.

4 a In the following two-node network, both of the nodes are M/M/1 queues with first-come first-served scheduling discipline. The service rate at node 1 is $\mu_1 = 2\gamma$ and the service rate at node 2 is $\mu_2 = 4\gamma$.



i) Give a condition on α for the network to be stable, i.e. for a steady-state probability distribution to exist.

ii) Assuming that the stability condition holds, determine the steady-state probability that there is 1 job at node 2.

iii) Suppose now that $\alpha = 0$ and $\gamma = 1$, and assume that the second node is replaced by a queue with Erlang-distributed service times. The Erlang distribution is obtained by summing $k = 4$ exponential random variables, each with rate $\lambda = 16$ job/s. Explain why Jackson's theorem does not hold with this model, but can still be analysed without the need for approximations. Then determine the mean queue-length at both queues.

b In a closed queueing network of M nodes, with paths existing between all pairs of nodes, and having population K, node m has service demand D_m $(1 \leq m \leq M)$. Assume that $D_1 = D_2 = \ldots = D_M$.

i) Express the normalizing constant, $g(K, M)$, as a function of K, M and D_1 only.

ii) Obtain an explicit formula for the network throughput, $X(K)$. Using this, determine the utilization of the nodes when K and M tend to infinity while keeping the ratio M/K constant.

iii) Assume now that $M = 2$, $K = 3$, $D_1 = D_2 = 3$. Compute the value of $g(K, M)$ using the convolution algorithm.

The two parts carry equal marks.