Computer Systems Modelling

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Computer Science Tripos, Part II

Michaelmas Term 2002

R. J. Gibbens

Problem sheet

William Gates Building
JJ Thomson Avenue
Cambridge
CB3 0FD

`http://www.cl.cam.ac.uk/`

Simulation

1. In what contexts is simulation an appropriate technique for performance evaluation? When it is inappropriate?

2. How would you generate random variates for the exponential distribution? Write pseudo code for a simulation component which models the arrival process of customers at a bank, given that the mean arrival rate is 40 customers per hour. What are the important events for this component (the generator).

3. How would you generate random variates from the mixed exponential distribution

$$f(x) = 1/3\lambda_1 \mathrm{e}^{-\lambda_1 x} + 2/3\lambda_2 \mathrm{e}^{-\lambda_2 x}$$

   where $\lambda_1 = 5, \lambda_2 = 10$, which describes the joint arrival process of two separate exponentially distributed streams of customers to a single queue?

4. Suppose that a simulation is constructed to estimate the mean response time, in milliseconds, of an interactive computer system. A number of repetitions are performed with the following results (measured in milliseconds): 4.1, 3.6, 3.1, 4.5, 3.8, 2.9, 3.4, 3.3, 2.8, 4.5, 4.9, 5.3, 1.9, 3.7, 3.2, 4.1, 5.1.

   Calculate the sample mean and variance of these results and hence derive a 95% confidence interval for the mean response time. You may wish to make use of the fact that $\mathbb{P}(Z > 1.96) > 0.025 > \mathbb{P}(Z > 1.97)$ where Z is a random variable with the unit normal distribution.

5. An experiment is performed to estimate the performance of a $M/M/1$ system with FIFO queueing. The response times of each of 1000 successive customers are recorded and the sample mean ($\overline{x}$) and sample variance (σ^2) of these numbers are calculated.

   A naïve student believes that $100(1-\alpha)\%$ confidence bounds for the mean can be derived by the expression

$$\overline{x} \pm \frac{z_{\alpha/2}\sigma}{\sqrt{1000}}$$

   where $z_{\alpha/2}$ is defined from the unit normal distribution in the usual manner.

   What mistake has this student made? Would their technique be valid if they were estimating the service time of this queue?

6. Using the inverse transform method show that

$$X = \left\lfloor \frac{\log(U)}{\log(1-p)} \right\rfloor + 1$$

   has a geometric distribution with parameter p where U has the $U(0,1)$ distribution.

7. Describe a procedure for the generation of the first T time units of a Poisson process of fixed rate λ.

8. Describe the variance reduction technique based on antithetic variables and give an example of how it is used.

9. For the variance reduction technique based on control variates derive the optimal choice of $c = c^*$ and its associated variance given in the lectures.

Operational analysis

1. In what kinds of context are the tools of operational analysis appropriate for performance evaluation? When are they inappropriate?

2. A printer and print queue are observed for 48 minutes during which time 16 jobs were completed. The mean queue length was 4 jobs. The printer could print 8 pages per minute. The mean job size was 12 pages.

   Assuming that the number of arrivals was equal to the number of completions, calculate

   (*a*) the throughput of the system in jobs-per-minute,

   (*b*) the mean residence time of a job,

   (*c*) the mean queuing time of a job, and

   (*d*) the utilization of the printer.

3. An queueing network described as *open* and *feed forward* is to be analyzed using operational laws. The mean arrival rate into the system is known to be λ.

   (*a*) Describe the significance of the two phrases in italics.

   (*b*) The mean service times, visit counts and connectivity are known for each device. Why is it still impossible, in general, to calculate an upper bound on the residence time of a job in the network?

   (*c*) Suppose, however, that there are never more than five jobs in the network at once. Outline how you could derive an upper bound on the residence time.

4. Suppose that busses run according to a regular timetable with a fixed inter-arrival time of 10 minutes. What is the mean waiting time of a customer arriving at the bus stop?

   Suppose instead that bus inter-arrival times are exponentially distributed, still with mean 10 minutes. What now is the mean waiting time of an arriving customer?

   Derive a general relationship between the mean waiting time and the first and second moments of the inter-arrival distribution. *Hint: if the inter-arrival time pdf is $f(x)$ then note that the probability of encountering an interval of length τ is proportional to both $f(\tau)$ and to τ.*

Queueing theory

1. In what situations does queuing theory provide appropriate techniques for performance evaluation? When does it not?

2. Show for the M/M/1 queue that the probability that there are n or more customers in the system is given by ρ^n.

   Use this result to find a service rate μ such that, for given λ, n, α where $0 < \alpha < 1$, the probability of n or more customers in the system is given by α.

   Find a value of μ for an M/M/1 queue for which the arrival rate is 10 customers per second, and subject to the requirement that the probability of 3 or more customers in the system is 0.05

3. What is the average length of each idle period of an M/M/1 server, given its arrival rate λ and service rate μ. What is the average length of a busy period?

4. Using the steady-state distribution of the number of jobs in a $M/M/1$ queueing system given in lectures derive the first and second moments of this distribution and hence the variance of the number of jobs present. Describe what happens as the load ρ increases.

5. Given an $M/M/1/K$ queue with $\lambda = 10$, $\mu = 12$ and $N = 15$ over what proportion of time are customers rejected from the queue? What is the effective arrival rate? What is the effective utilization of the server?

6. Repairing a computer takes 4 stages in sequence, namely removing the lid, finding the faulty part, replacing it, and reassembling the machine. Each step is independent and exponentially distributed with mean 3 minutes. What is the coefficient of variation of the repair time? Construct a Markov chain model of this system, assuming an infinite population of machines.

7. A closed queueing network (shown below) comprises two $M/M/1$ nodes, **A** and **B** between which n identical jobs circulate. The nodes have service rates μ_A and μ_B respectively. Upon completion at **A**, a job moves to **B** with probability p and otherwise it remains at **A**. Similarly, upon completion at **B**, a job moves to **A** with probability q.

   Derive a Markov-chain model of this system, explaining what each state in the model signifies and what transition rates exist between states.

Past exam questions

1. **1995 Paper 8 Q11.** Given that in a balanced system with K devices and N customers, the utilisation of each device is given by

$$U = \frac{N}{N + K - 1}$$

derive a formula for the response time in terms of throughput, the number of devices and the average service demand at each device. [6 marks]

A system consists of three types of devices, A, B and C. Customers require service at each type of device but do not care at which particular device they are served. The numbers of each type of device and average service requirements per customer are

	number of devices	average service demand
A	48	48 ms
B	24	24 ms
C	18	18 ms

so that, for example, a customer requires on average 48 ms of service at a type A device.

Give bounds for the system response time at a throughput of 500 customers per second when a scheduling policy ensures that

(a) no device is more than 1.5 times as busy as the average for devices of the same type

(b) no device is more than 1.8 times as busy as the average for devices of the same type

[9 marks]

What can you say about the response time if no limit on utilisation skew across devices of the same type is guaranteed? [5 marks]

2. **1995 Paper 9 Q11.** In queueing networks, what is meant by a *closed* system? [4 marks]

Consider two closed systems. One has two devices, A and B, and three customers, the other three devices, A, B and C, and two customers. Both have exponentially distributed service times which are device dependent but customer independent. In the first system a customer completing service at a device always moves to the other device. In the second system a customer completing service moves to one of the other two devices with equal probability.

Draw state diagrams for the Markov chains representing these systems. Choose one system to solve for device utilisation in terms of service rates. [10 marks]

For the chosen system, when the service rates are equal does the utilisation of each device correspond to that for a balanced system ($U = \frac{N}{N+K-1}$ where N is the number of customers and K the number of devices)? [3 marks]

Describe the state space for a Markov chain for one of the systems if the service rates were both customer and server dependent. [3 marks]

3. **1996 Paper 8 Q3.** A telephone exchange multiplexes 64Kb/s voice calls onto a 256Kb/s trunk line. New calls have an exponentially distributed inter-arrival process, with a mean of 20 seconds, and the call holding time is exponentially distributed with a mean of 60 seconds.

(a) Draw a diagram of a Markov Chain which models the system, labelling the state transitions with their rates where appropriate. What is the necessary condition for stability of this system ? [5 marks]

(b) Derive an expression for the probability that an arriving call finds k calls in progress, for $k \geq 0$, and thence calculate the probability that a caller finds the exchange engaged, given the parameters above. [15 marks]

4. **1996 Paper 9 Q3.** The Erlangian distribution E_r, with parameters (r, μ) is given by

$$f_X(x) = \frac{\lambda(\lambda x)^{r-1}}{(r-1)!}\mathrm{e}^{-\lambda x} \qquad x \geq 0$$

(a) Given an $M/E_2/1$ queue, draw a Markov Chain describing the queueing system, and derive a formula for the variance of the service time distribution. Give an example of a queueing system in which the use of an Erlang service time distribution would be useful. [5 marks]

(b) Describe how random variables from a given distribution function $f_X(x)$ can be sampled for use in a discrete event simulator. [3 marks]

(c) Use your answer from the previous question to develop pseudo-code for a function in a discrete event simulator which when called returns a sampled value from the distribution function E_r. State any assumptions made and explain any arguments to the function. Comment on the efficiency of your code. [12 marks]

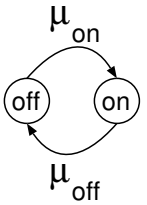
5. **1997 Paper 8 Q3.** A group of smart computer scientists decide to earn some money in the vacation by buying a punt and running a chaueured punt tour operation on the Cam. Based at Magdalene Bridge, they run tours upriver showing tourists the delights of the Backs. As those familiar with punt trips will know, journeys take an exponentially distributed time to complete, particularly when accompanied by strawberries and champagne. In this case our punt operators have measured that the average trip takes 30 minutes. Tourists (being tourists) behave rather randomly and arrive at the quay independently, according to Poisson distribution with mean 10 per hour. Each punt can accommodate six tourists, but departs as soon as there is at least one tourist wanting to take the tour.

(a) Draw a Markov chain model for the punt tour operation, annotating it with the appropriate rates and state information. Briefly explain the diagram. [5 marks]

(b) Propose a model which can safely be used as a worst-case approximation to the system above. Why is your model conservative? Using your new model, calculate the expected length of the queue of tourists. What constraints are required to ensure that the queuing system remains stable? [5 marks]

(c) Now assume that tourists do not join the queue if there are already six others waiting. If each tourist pays 5 per trip, and a punt costs 10 per hour to run, calculate the expected hourly prot of the punt company. Propose a scheme which would enable the entrepreneurs to increase their prots, but do not solve a model of the new system. What is the utilisation of the punts in the old and new schemes? [10 marks]

6. **1997 Paper 9 Q3.** The two-state Markovian on-o source is used in computer networking research as a model of bursty network traffic. The source is pictured in the figure below. When in the on state it generates fixed-sized network packets at a *constant* rate of 100 packets/second. Each set of packets generated in the on state is called a *burst*. When in the off state the source is silent. The residence time in each state is exponentially distributed. The *burstiness* of the source is dened as the ratio of its peak and mean rates.



A traffic modeller wants to write a simulation module which emulates this source, with the requirement that it transmits bursts of traffic with a mean length of 25 packets, and a burstiness of 20.

(*a*) Calculate the rates μ_{off} and μ_{on}. [10 marks]

(*b*) Briefly describe how to generate values which are distributed exponentially for use in the simulator. [3 marks]

(*c*) State Little's Law. Outline a proof of Little's Law with the aid of a diagram. [7 marks]

7. **1998 Paper 8 Q3.** Define the term *Markov Chain*. Why is the Markov property useful in modelling queueing systems? [5 marks]

   Consider a birth-death queueing system with the following birth and death coefficients in which the state index represents the number of customers in the system:

$$\begin{aligned} \lambda_k &= (k+2)\lambda \quad k = 0, 1, 2 \ldots \\ \mu_k &= k\mu \qquad\quad k = 1, 2 \ldots \end{aligned}$$

   All other coefficients are zero. Solve for p_k, the set of equilibrium probabilities for all states k, for $k = 0, 1, 2 \ldots$ State how you would find the average number of customers in the system. [15 marks]

8. **1998 Paper 9 Q3.** An $M/M/m$ queue has an arrival process with mean rate λ, and processes customers at a mean rate of μ.

   (*a*) What are the distributions and parameters of the inter-arrival and service times of customers? [3 marks]

   (*b*) Sketch an outline proof showing that the distribution of the departure process from the queue is the same as that of the arrivals process. [10 marks]

   Briefly contrast analytical queueing analysis and discrete event simulation with regard to their fields of applicability and other important considerations for the systems modeller. [7 marks]

9. **2000 Paper 7 Q8.** What is meant by the term *memoryless* as used in describing a Markov chain? [3 marks]

   What limitation does this place on using Markov chains to model real systems? [3 marks]

   "As systems become saturated their response time becomes unpredictable."
   Why is this? Illustrate your answer using an M/M/1 queueing system. [10 marks]

   Show, by drawing the state transition diagram of a Markov chain, how arrival processes with inter-arrival times that are not exponentially distributed can be modelled. [4 marks]

10. **2000 Paper 9 Q7.** A database system has a central processor and three (different) discs.

    Measurements are taken for 1000 transactions on a lightly loaded system and the following observations are made.

- The CPU scheduler initiated or resumed transaction processing 10,000 times. The total CPU usage was 25 seconds.
- Disc 1 made 5000 transfers with an average transfer time of 10 ms.
- Disc 2 made 2000 transfers with an average transfer time of 50 ms.
- Disc 3 made 2000 transfers with an average transfer time of 20 ms.

Derive the visit counts, service times and transaction service demands. What is the bottleneck device? What is the maximum throughput of the system measured in transactions per second? [6 marks]

Describe *two* balanced systems which bound the throughput of the system. What is the maximum throughput of these systems? [7 marks]

Recall that the throughput of a balanced system with K devices, N customers and service demand D per device is

$$X(N) = \frac{N}{(N+K-1)} \times \frac{1}{D}$$

How many transactions do you expect to be in the system with a throughput of 7 transactions per second? [7 marks]

11. **2001 Paper 7 Q8.** What criteria would you consider when selecting between a model based on queueing theory and one based on simulation? When might you use both approaches? [5 marks]

    Describe the structure of a *discrete event simulator*. What is the principal data structure involved? [5 marks]

    A queueing network is characterised by a set of *visit counts*, V_i, and *per-visit service requirements*, S_i, for each of N devices. Derive upper bounds on the system throughput (i) when the load is very low and (ii) as the load tends to infinity. [5 marks]

    In what situations may the bounds be particularly imprecise? What can be done to construct tighter bounds for the system throughput? [5 marks]

12. **2001 Paper 8 Q14.** Two servers operate with different performance characteristics at mean rates μ_1 and μ_2. You wish to combine them into a single system by associating each server with a separate FIFO queue and dispatching incoming work items to the first queue with probability p_1 and to the other queue with probability p_2. Incoming items arrive at a rate λ and none are discarded from the system.

    You may assume that the inter-arrival-time distribution and both service-time distributions are exponential, that there is no limit on the queue lengths and that the population size is infinite.

    Using Kendall notation, describe the first server and its queue. Construct a Markov-chain model for this part of the system. [2 marks]

    Let $q_{k,i}$ denote the probability that there are exactly i items of work in server k and its queue. By using detailed flow balance equations or otherwise express $q_{k,i}$ in terms of λ, p_k and μ_k. [6 marks]

    Hence derive T_k, the mean response time of work items served at k. [6 marks]

    Suppose that the system administrator wishes to ensure that work items receive the same mean response time irrespective of which server they visit. Express p_1 in terms of λ, μ_1 and μ_2. Qualitatively, when is it reasonable to consider dispatching work to both servers to maintain an equal mean response time? How will the system behave at other times? [6 marks]

13. **2002 Paper 7 Q8.** Consider an $M/M/1$ queue and represent the state of the queue by the number of customers present.

(a) Draw a state diagram for the Markov chain describing the state of the queue showing the possible states and transition rates. Briefly explain the diagram. [4 marks]

(b) What is the condition for the existence of a steady-state equilibrium distribution p_k $(k = 0, 1, 2, \ldots)$ for the number present? [2 marks]

(c) Determine the steady-state average number of customers present. [4 marks]

(d) Use Little's law to determine the steady-state average response time that a customer spends in the system. [2 marks]

(e) Suppose that a single communication channel is used to carry data items sent by various sources connected to the channel. Assume that each source generates a stream of data items with inter-arrival times which are exponential at rate 2 items/second and that all sources are statistically independent. All the items wait in a single queue and are transmitted one at a time. The transmission times are exponentially distributed with mean 25 milliseconds and are statistically independent. Determine the largest number of sources that can be connected to the channel according to each of the following two criteria:

(i) The channel is not saturated.

(ii) The steady-state average response time for an item must not exceed 100 milliseconds.

[8 marks]

14. **2002 Paper 8 Q14.** Consider an $M/G/1$ queue.

(a) What is the condition for the non-saturation of the server? [2 marks]

(b) State the *Pollaczek-Khintchine* formula for the steady-state average number of customers present in the system. [4 marks]

(c) Suppose that Edward and Ursula are two applicants for a vacancy as a bank teller. It has been determined by empirical testing that Edward's service times are exponentially distributed with mean 0.9 minutes, while Ursula's are uniformly distributed between 0.8 and 1.2 minutes. It is known that customers arrive at the bank teller's window with inter-arrival times which are exponentially distributed and at the rate of 50 per hour.

(i) Can both applicants cope with the load?

(ii) If so, which one should be employed so as to minimize the steady-state average number of customers present?

In both cases, justify your answer. [8 marks]

(d) Given a sequence of pseudo-random numbers $U_1, U_2, \ldots$ distributed uniformly between 0 and 1 explain briefly how to construct pseudo-random sequences for the inter-arrival times of bank customers and for the service times of both Edward and Ursula. [3 marks]

(e) Briefly compare the advantages and disadvantages of the analytical queueing theory and the discrete event simulation approaches to determining performance measures by the above bank employer. [3 marks]