COMP9334 Capacity Planning of Computer Systems and Networks
Assignment (Version 1.0), Term 1, 2022

Due 5:00pm, Fri 18 March 2022 (Week 5)

Change log and version info

Updates, changes and clarifications will appear in this box.

- Version 1.0 issued on 28 February 2022

Instructions

(1) There are 3 questions in this assignment. Answer all questions.

(2) The total mark for this assignment is 20 marks.

(3) In answering the questions, it is important for you to show your intermediate steps and state what arguments you have made to obtain the results. You need to note that both the intermediate steps and the arguments carry marks. Please note that we are **not** just interested in whether you can get the final numerical answer right, we are **more** interested to find out whether you understand the subject matter. We do that by looking at your intermediate steps and the arguments that you have made to obtain the answer. Thus, if you can show us the perfect intermediate steps and the in-between arguments but get the numerical values wrong for some reason, we will still award you marks for having understood the subject matter.

If you use a computer program to perform any part of your work, you **must** submit the program or you lose marks for the steps.

(4) The submission deadline is 5:00pm Friday 18 March 2022. Late submission will cap the maximum mark that you receive. Submissions after 5:00pm on Sunday 20 March 2022 will no longer be accepted.

(5) Your submission should consist of:

(a) A report describing the solution to the problems. This report can be typewritten or a scan of handwritten pages. This report must be in pdf format and must be named assignment.pdf. The submission system will only accept the name assignment.pdf.

(b) One or more computer programs if you use them to solve the problems numerically. You should use zip to archive all the computer programs into one file with the name supp.zip. The submission system will only accept this name. The report must refer to the programs so that we know which program is used for which part.

(6) Submission can be made via the course website.

(7) You can submit as many times as you wish before the deadline. A later submission will over-write the earlier one.

Question 1 (3 marks)

An interactive computer system consists of a CPU and three disks. We will use disk-1, disk-2 and disk-3 to refer to these three disks. The system was monitored for 60 minutes and the following data were available:

T : monitor time = 60 min × 60s = 3600 s

Number of completed requests by the system	789 → X(0)
Visit ratio of disk-1	4.5
Visit ratio of disk-2	5.9
Visit ratio of disk-3	5.1
Visit ratio of CPU	25.7
Busy time of disk-1	2917 seconds B(disk-1)
Busy time of disk-2	2718 seconds B(disk-2)
Busy time of disk-3	2867 seconds B(disk-3)
Busy time of the CPU	2665 seconds B(CPU)

Answer the following questions.

① $U = \frac{B}{T}$ ② $X_{(0)} = \frac{C}{T}$ ③ $D_{(j)} = \frac{U_{(j)}}{X_{(0)}}$

(a) Determine the service demands of disk-1, disk-2, disk-3 and the CPU.

(b) Use bottleneck analysis to determine the asymptotic bound on the system throughput when there are 4 interactive users and the think time is 20 seconds.

Reminder: If you use a computer program to derive your numerical answers, you must include your computer program in your submission. Do not forget to show us your steps to obtain your answer.

$$X(0) \leq \min\left[\frac{1}{\text{Max}(D_j)}, \frac{N}{\sum_{i=1}^{K} D_i + \text{Thinky Time}}\right]$$

Question 2 (7 marks)

A call centre has 2 staff to deal with customer enquires. The centre has a dispatcher to direct the calls automatically to one of the staff. The dispatcher does not contain any queueing facilities. At each staff's terminal, there is a facility to queue up to 3 calls. The queueing network at the call centre is depicted in Figure 1.

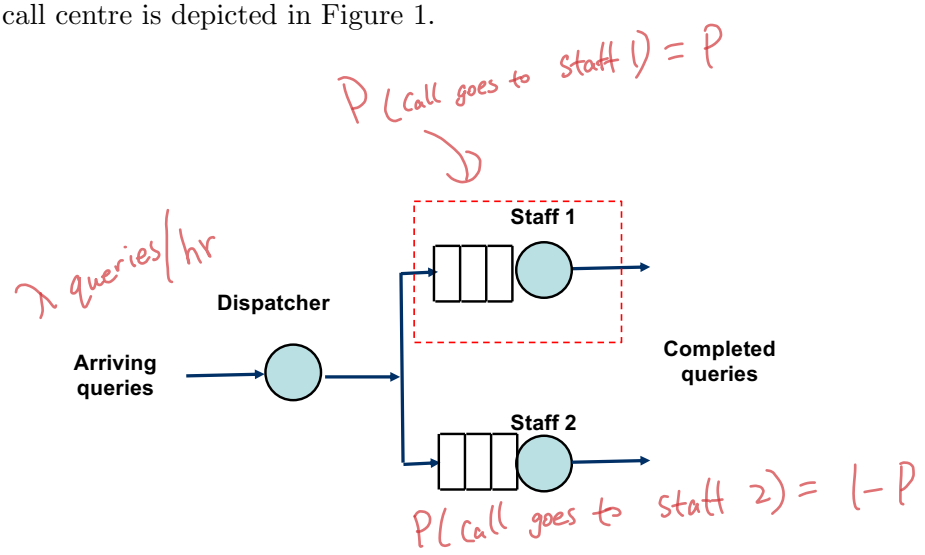


Figure 1: Depiction of the call centre.

The centre receives on average λ queries per hour. The arrivals can be modelled by using the Poisson distribution.

When a query arrives at the dispatcher, it will send the query to Staff 1 with a probability of p and to Staff 2 with a probability of $1-p$. Note that the dispatcher does not communicate with the staff's terminals, so it is possible that the dispatcher sends a query to a terminal that has a full queue. You can assume that the dispatcher takes a negligible time to perform its work and no queries will be dropped at the dispatcher.

Staff 1 and Staff 2 can complete, respectively, on average μ_1 and μ_2 queries per hour. The amount of time required by each staff is exponentially distributed.

When a query arrives at a staff's terminal, it will be answered straight away if the staff is not busy. Otherwise, the terminal will place the call in its queue if the queue is not full. If the call arrives when the queue is full, then the call is rejected.

Answer the following questions:

(a) Formulate a continuous-time Markov chain for the part of the call centre consisting of Staff 1 and their three waiting slots, i.e. the part enclosed by the red dashed lines in

Figure 1. Your formulation should include the definition of the states and the transition rates between states. The transition rates should be expressed in terms p, λ and μ_1.

(b) Write down the balance equations for the continuous-time Markov chain that you have formulated.

(c) Derive the expressions for the steady state probabilities of the continuous-time Markov chain that you have formulated.

(d) Assuming that $p = 0.4$, $\lambda = 5.7$ and $\mu_1 = 6.1$. Determine the probability that a query that is dispatched to Staff 1 will be rejected.

(e) Assuming that $p = 0.4$, $\lambda = 5.7$, $\mu_1 = 6.1$ and $\mu_2 = 6.5$, determine the mean waiting time of the queries that have not been rejected by the call centre. Note that Part (d) considers only queries that have been dispatched to Staff 1 but Part (e) considers the whole call centre.

Staff 1: $P(0) + P(1) + P(2) + P(3)$

Staff 2: $P(0) + P(1) + P(2) + P(3)$

Staff 1 + 2

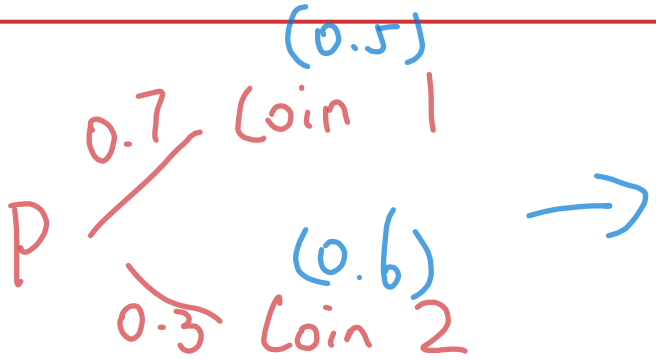
W3A P22

Hint:

- There is a mistake that some people may make regarding the calculation of the mean waiting time in Part (e). We will not tell you exactly what the mistake is but the following example of probability calculations will illustrate that. Let us assumed that you have two coins, which we will refer to as Coin 1 and Coin 2. Coin 1 is a fair coin and the mean number of heads you get is 0.5. Coin 2 is a biased coin and the mean number of heads you can get is 0.6. Let us say you do the following:
  - You randomly pick one of the two coins with the probabilities of picking Coins 1 and 2 being, respectively, 0.7 and 0.3. You toss the coin picked. You repeat this many times.

  You want to calculate the mean number of heads that you will get. A wrong answer is 0.55. The correct answer should be 0.53. △△△

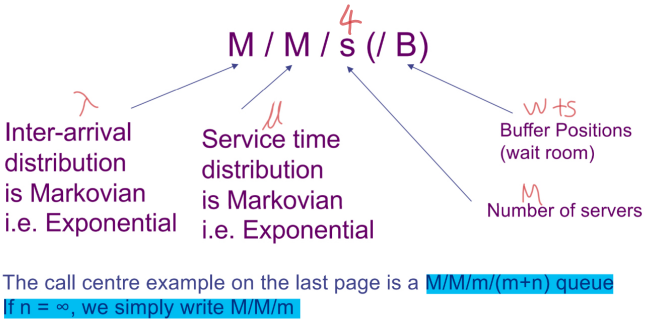
Reminder: If you use a computer program to derive your numerical answers, you must include your computer program in your submission. Do not forget to show us your steps to obtain your answer.



$(0.7 \times 0.5) + (0.3 \times 0.6)$

$= 0.35 + 0.18$

$= 0.53$



Question 3 (10 marks)

This question is based on the system illustrated in Figure 2a. The system consists of a dispatcher at the front-end and n servers at the back-end. We will use the value of $n = 4$ for explanation in Figures 2-4 but you will need to vary the value of n later on when answering the questions. Note that this system has no queueing facilities at neither the dispatcher nor the servers.

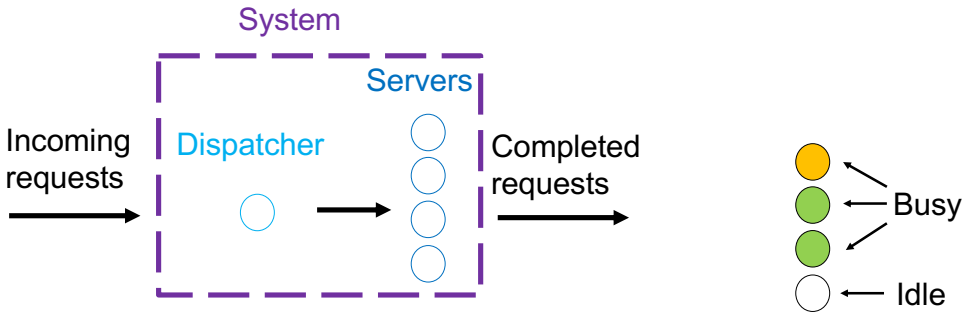
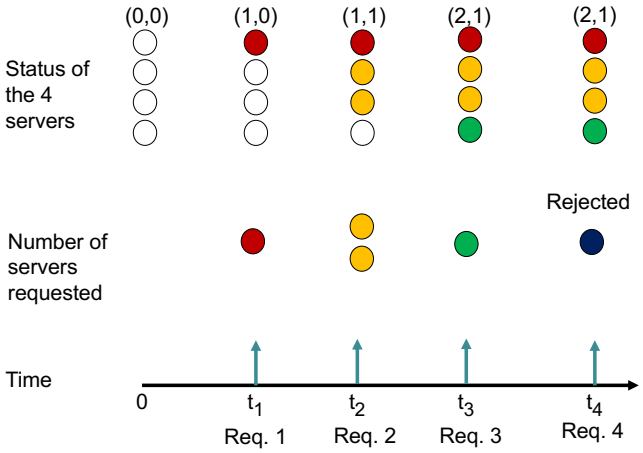


Figure 2: (a) Pictorial representation of the system. (b) Conventions in Figures 3a and 3b: an unfilled circle means the server is idle and a filled circle means the server is idle.

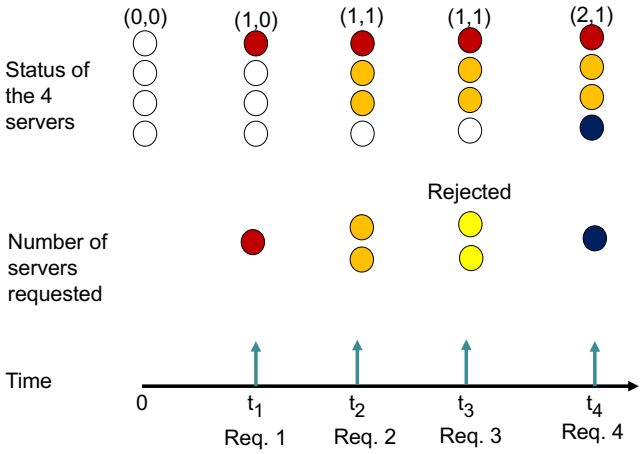
This system is used to serve two classes of requests, which we will refer to as Class 1 and Class 2. Each request from Class 1 requires 1 server for processing while each request for Class 2 requires 2 servers for processing.

A function of the dispatcher is to decide whether an arriving request can be admitted or otherwise rejected. If there is at least an idle server at the time that a Class 1 request arrives, then the request will be admitted and dispatched to any one of the idle servers. If there are at least 2 idle servers at the time that a Class 2 request arrives, then the request will be admitted and dispatched to any two of the idle servers. If there are insufficient idle servers to admit a request, then that request will be rejected.

Figures 3a and 3b illustrates the admission and rejection of requests. These illustrations assume that all the servers are idle at time 0. Also, we assume that the processing of the admitted requests will not have been completed by the arrival time of Request 4, which means there are no departures in the time period illustrated in these two figures. The number of servers wanted by each request is shown. If there are sufficient number of idle servers available at the time the request arrives, the request will be admitted. Note that for the server status, an unfilled circle means a server is idle and a filled circle means a server is busy, see Figure 2b for the convention. Note that the server status shown is for the time after the admission decision of an arriving request has been made. As an example, in Figure 3a, Request 4 is rejected because there are no idle servers at the time that this request arrives. As another example, in Figure 3b, Request 3 is rejected because there is only one idle serve available at the time that this request arrives but this request wants two servers.



(a) Example 1.



(b) Example 2.

Figure 3: Illustrating the admission and rejection decisions.

You can assume that the dispatcher takes a negligible time to perform its function. This means that if a request is admitted, then its processing begins at its time of arrival because the dispatcher takes negligible time to admit a request and send it to the server(s).

For each request from Class 1, the amount of work needed to perform by a server is exponentially distributed with mean w_1. We use Figure 4 to explain the processing requirements of a Class 2 request. Figure 4 shows that, if a Class 2 request is admitted, then both servers allocated to process that request will start the processing at the same time and both servers will also complete the processing at the same time. The amount of work for Class 2 requests will be specified on a *per server* basis. We will assume that, for each request from Class 2, the amount of work needed to perform per server is exponentially distributed with mean w_2. Note that requests will leave the system once their processing is complete.

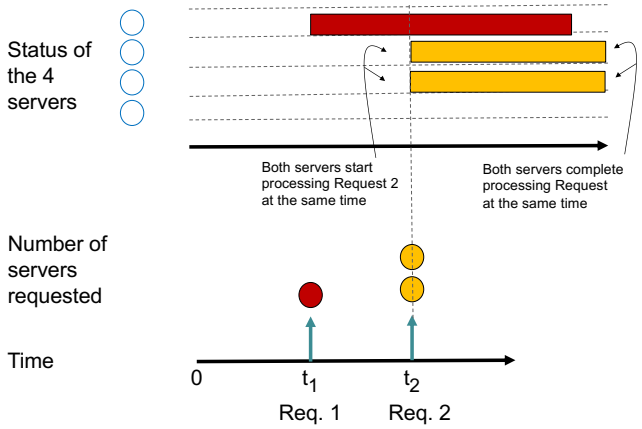


Figure 4: The processing of a Class 2 request will start at the two allocated servers at the same time and also end at the same time.

The mean arrival rates of Class 1 and Class 2 requests are respectively, λ_1 and λ_2; and their arrival distributions are Poisson. Furthermore, you can assume that the arrivals from the two Classes are independent. The units for both λ_1 and λ_2 are number of requests per hour.

Each server has a constant processing rate of μ. The unit of μ is amount of work per hour. The units for w_1 and w_2 are the amount of work. You can assume that there is no processing overhead so the amount of time needed for processing is the amount of work divided by the processing rate.

The system considered in this question can be modelled by a continuous-time Markov chain whose state is the tuple (r_1, r_2) where r_1 and r_2 are, respectively, the numbers of Class 1 and Class 2 requests in the servers. For example, in Figure 3a, the state of the system is $(2, 1)$ after Request 3 has been admitted. See Figures 3a and 3b for other examples.

Answer the following questions. Please note that you will be using a specific value of n for Parts (a) and (b), but you will need to vary n later on in Part (c).

(a) Assuming that $n = 4$, formulate a continuous-time Markov chain for the system using the state definition given earlier. You can answer this question by drawing a state transition diagram with all the states and transitions. You can express the transition rates in terms of λ_1, λ_2, w_1, w_2 and μ.

(b) Assuming that $n = 4$, $\lambda_1 = 2.7$, $\lambda_2 = 1.5$, $w_1 = 10.4$, $w_2 = 15.3$ and $\mu = 70$. Answer the following questions.

(i) What are the steady state probabilities of the states for the continuous-time Markov chain?

(ii) Determine the probability that an arriving Class 1 request will be rejected.

(iii) Determine the probability that an arriving Class 2 request will be rejected.

(iv) Determine the probability that an arriving request will be rejected. Note that the hint in Question 2 is applicable.

(c) Assuming that $\lambda_1 = 2.7$, $\lambda_2 = 1.5$, $w_1 = 10.4$, $w_2 = 15.3$ and $\mu = 70$. What is the smallest value of n that can reduce the probability of rejecting an arriving request to a level lower than 0.05?

Remark: We mention in the lectures that the concept of a server in Queueing Theory is very general. In the context of this question, you can think about a server as a certain amount of bandwidth. If a request requires a higher number of servers, then this request needs more bandwidth. So, a Class 1 request may be an audio call which needs less bandwidth and a Class 2 may be a video call which needs more bandwidth.

Reminder: If you use a computer program to derive your numerical answers, you must include your computer program in your submission. Do not forget to show us your steps to obtain your answer.

– – –– End of assignment – – ––