COMP9334 Capacity Planning
Assignment, Term 1, 2019

4 March 2019 (Version 1.1)

Change log

Note: New text is shown in red coloured font. Deleted text is retained and shown as strikethrough, e.g. ~~this is deleted text~~.

- Version 1.1. (4 March 2019) Error in the observation time in Question 1.

Instructions

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

(2) The total mark for this assignment is 15 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 11:00pm Monday 25 March 2019. Late submission will cap the maximum mark that you receive. Submissions after 11:00pm on 27 March 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 tar/zip/rar to archive all the computer programs into one file with the name supp.tar, supp.zip or supp.rar. The submission system will only accept these names. 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 overwrite the earlier one.

Question 1 (4 marks)

An interactive computer system consists of a CPU and a disk. The system was monitored for ~~60~~ **90** minutes and the following measurements were taken:

Number of completed jobs	676
Number of CPU accesses	1,377
Number of disk accesses	1,515
CPU busy time	4,729 seconds
Disk busy time	2,565 seconds

Answer the following questions.

(a) Determine the service demand for each device of the system.

(b) In the lecture, we told you that you could identify the bottleneck of the system by using service demand. For the setting of this question, do you think it is possible to determine the bottleneck of the system without calculating the service demands? Justify your answer.

(c) Use bottleneck analysis to determine the asymptotic bound on the system throughput when there are 30 interactive users and the think time per job is 31 seconds.

(d) Using your results in Part (c), compute the minimum possible response time when the number of interactive users is 30.

Note: 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.

Question 2 (4 marks)

This question is based on the server farm in Figure 1. The server farm consists of a dispatcher and two computer systems, which are labelled as Systems 1 and 2. The purpose of the dispatcher is to route the incoming requests to either of the two systems. The policy of the dispatcher is to route an incoming request to System 1 with a probability of p, and to System 2 with a probability of $1 - p$. You can make the following assumptions.

- The arrivals to the dispatcher is Poisson distributed with a mean arrival rate of λ requests/s where $\lambda = 20$
- The processing rate of System 1 is μ_1 requests/s where μ_1 = 10
- The processing rate of System 2 is μ_2 requests/s where μ_2 = 15
- The dispatcher takes a negligible time to process a request. The transmission of a request from the dispatcher to the chosen system also takes negligible time.
- Both Systems 1 and 2 have their own queue. You can assume infinite buffer space for both systems.
- Requests will leave the systems, hence the server farm, once they have been completed.
- The response time of the server farm is dominated by the response time of the two systems.

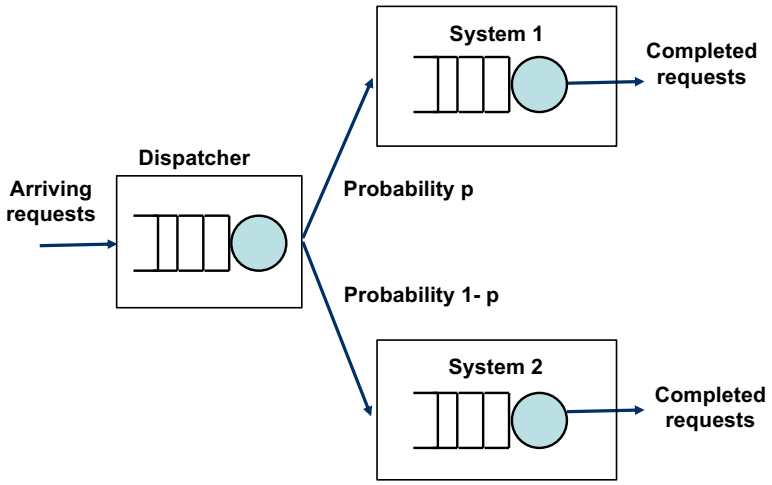

Figure 1: Figure for Question 2.

Note that this question's server farm has two systems which have different processing rates. In general, it is common for server farms to have computer systems of heterogeneous speeds. This is because of incremental expansion of a server farm where the computer systems were purchased at different times. This explains why we chose different processing rates for the two systems. For the

server farm in this question, if you choose $p = 0.5$ then the performance of the server farm will be very bad. You want to choose a value of p to balance the load. This is known as the load balancing problem in performance analysis.

Answer the following questions:

(a) Find the probability p so that Systems 1 and 2 have the same utilisation.

(b) Determine the mean response time of the server farm for the value of p that you have calculated in (a).

(c) Determine the value of p so that the mean response time of the server farm is the smallest possible.

Notes:

- You can do part (c) analytically, graphically or numerically using a computer program.
- There is a mistake that some people may make regarding the calculation of the mean response time of the server farm. 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.
- For the specific numerical values of λ, μ_1 and μ_2 used earlier, you will find that if you balance the utilisations of the two systems, the resulting mean response time (which is your answer to part (b)) is pretty close to the minimum response time you find in part (c). An interesting question is whether it is true in general that balancing the utilisation will lead to close to minimum mean response time. By in general, we mean whether it holds for other values of λ, μ_1 and μ_2. I will leave you to investigate that as an open problem.

Note: 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.

Question 3 (7 marks)

This question is concerned with the reliability of a data centre.

A data centre has 4 machines. For each machine, the time-to-next-failure is exponentially distributed with a mean of 600 minutes. The failure of a machine is independent of the others.

The data centre has a repair team consisting of a team leader and a trainee. The time required by a repair staff to repair a machine is exponentially distributed. The mean time taken by the team leader to repair a machine is 60 minutes. However, the trainee takes on average 90 minutes to repair a machine.

The mode of operation of the repair team is:

- Each failed machine will be repaired by exactly one staff from the repair team.
- Once a staff has started repairing a machine, that staff will continue to work on it until completion.
- If at a certain point in time, the status of the data centre goes from all machines working to 1 machine failed, then the trainee will work on repairing that failed machine. For other failure scenarios, a newly failed machine will be repaired by any repair staff who is available.

The repair cycle of the data centre can be modelled by a Markov chain whose state is a 3-tuple (n, ℓ, t) where

- n is the number of machines that have failed
- ℓ is either 0 or 1. ℓ is 1 if the team leader is busy, otherwise it is 0.
- t is either 0 or 1. t is 1 if the trainee is busy, otherwise it is 0.

Answer the following questions:

(a) Derive the state transition diagram for the Markov chain that describes the above problem. The diagram needs to include both the states and the transition rates. Explain how you arrive at your state transition diagram.

(b) Derive the state balance equations for the Markov chain.

(c) Determine the steady state probability of all the states of the Markov chain.

(d) Compute the probability that at least three machines are available.

(e) Compute the mean number failed machines.

(f) Compute the mean-time-to-repair (MTTR) for this data centre.

Note: 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 $---$