COMP9334 Capacity Planning of Computer Systems and Networks
Assignment (Version 1.0), Term 1, 2021

Due 5:00pm, Fri 19 March 2021

Change log and version info

Updates, changes and clarifications will appear in this box.

- Version 1.0 issued on 1 March 2021

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 19 March 2021. Late submission will cap the maximum mark that you receive. Submissions after 5:00pm on Sunday 21 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 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 dual-core CPU and a disk. We will use core-1 and core-2 to refer to the two cores of the CPU. The system was monitored for 60 minutes and the following measurements were taken:

Number of completed jobs	1347
Number of accesses to core-1	2087
Number of accesses to core-2	2348
Number of disk accesses	2412
Busy time of core-1	2828 seconds
Busy time of core-2	1728 seconds
Disk busy time	2665 seconds

Answer the following questions.

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

(b) 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 15 seconds.

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 (7 marks)

A call centre has 3 staff to deal with customer enquires. The centre has an automatic dispatcher to direct the calls to the staff. The dispatcher has a queue that can hold up to 2 calls but there are no queueing facilities at the staff's terminals. The queueing network at the support centre is depicted in Figure 1.

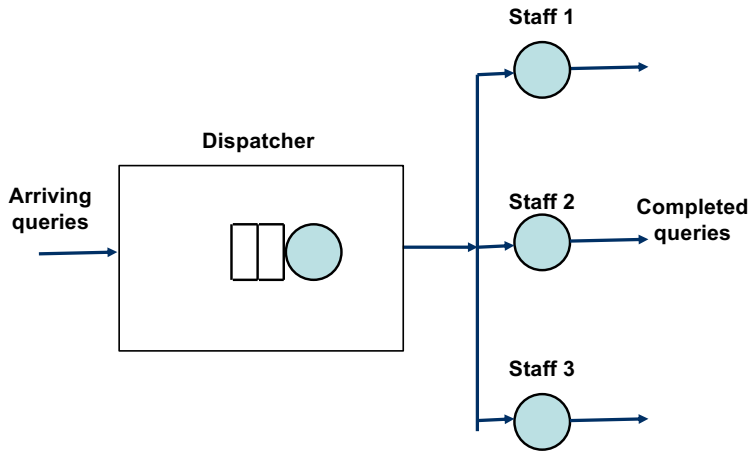

Figure 1: Figure for Question 2.

The centre receives on average 12.7 queries per hour. The arrivals can be modelled by using the Poisson distribution.

Each staff can complete on average 4.1 queries per hour. The amount of time required by each query is exponentially distributed.

When a query arrives at the dispatcher, it will accept the query if the dispatcher queue is not full, otherwise the query will be rejected. If a query is accepted and the queue is not empty, the query will be placed at the end of the queue. If a query is accepted and the queue is empty, then the query will be placed in the queue if all staff are busy, otherwise it will be sent to an idling staff. A query will leave the system after its processing is completed. Whenever a staff becomes idle, he/she will take the query from the front of the queue if there is one.

Answer the following questions:

(a) Formulate a continuous-time Markov chain for a system described above with 3 staff and 2 waiting slots. Your formulation should include the definition of the states and the transition rates between states.

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

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

(d) Determine the probability that an arriving query will be rejected.

(e) Determine the mean waiting time of an accepted query in the queue.

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 (10 marks)

This question is based on the server farm in Figure 2. The server farm consists of a dispatcher and two computer systems, which are labelled as Systems 1 and 2. Modern day server farms typically consist of systems of heterogeneous hardware specifications. This is due to incremental expansion where the computer systems are purchased at different times. In this question, we will assume that System 1 has a lower processing rate than System 2.

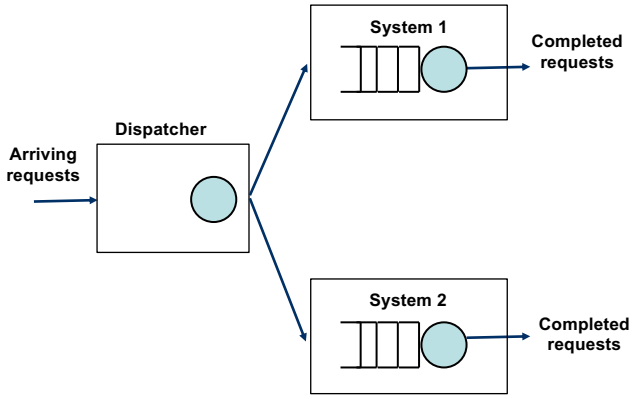

Figure 2: Figure for Question 3.

As depicted in Figure 2, the dispatcher does not have any queues. The maximum number of jobs in Systems 1 and 2 are, respectively, n_1 and n_2. Since each system has a single server, the number of waiting slots in the two systems are $n_1 - 1$ and $n_2 - 1$. We will specify what n_1 and n_2 are later on.

The purpose of the dispatcher is to route an incoming request to either of the two systems or to reject the request. The dispatcher uses a policy to decide how to route an incoming request. Let us assume that at the time that a request arrives, the number of jobs in Systems 1 and 2 are, respectively, j_1 and j_2. The dispatcher policy, in algorithmic format, is:

Algorithm 1: The dispatcher routing policy

```
   Data: j1,j2,n1,n2
1  if j1 < j2 and j1 < n1 then
2  |  Route the incoming request to System 1;
3  else
4  |  if j2 < n2 then
5  |  |  Route the incoming request to System 2;
6  |  else
7  |  |  Reject the incoming request;
8  |  end
9  end
```

The policy has been designed to make use of the faster system (i.e. System 2) before using the slower one in order to reduce the mean response time of the server farm. The policy achieves that by choosing System 1 if there are fewer jobs in System 1 than System 2 (i.e. the comparison $j_1 < j_2$ in Line 1) and the queue in System 1 is not full (i.e. $j_1 < n_1$ in Line 1). If an incoming request will not be sent to System 1, then it be sent to System 2 provided that the queue in System 2 is not full (i.e. $j_2 < n_2$ in Line 4). An incoming request will be rejected if it cannot be sent to neither system.

For this question, you can assume the following:

- The arrivals to the dispatcher are Poisson distributed with a mean arrival rate of λ requests/s where $\lambda = 1$.
- The service time distribution in both servers is exponentially distributed.
- The processing rate of Server 1 is μ_1 requests/s where $\mu_1 = 0.5$.
- The processing rate of Server 2 is μ_2 requests/s where $\mu_2 = 0.7$.
- The value of n_2 is 5.
- The dispatcher takes a negligible time to process a request and makes a routing decision. The transmission of a request from the dispatcher to the chosen server also takes negligible time. Overall, these assumptions imply that the response time of the server farm is dominated by the response times of the two systems.
- Requests will leave the server farm, once they have been completed.

Answer the following questions:

(a) Assuming that $n_1 = 1$, answer the following questions:

(i) Formulate a continuous-time Markov chain for the server farm. Your formulation should include the definition of the states and the transition rates between states.

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

(iii) Determine the mean response time of System 1.

(iv) Determine the mean response time of the server farm.

(b) Determine the response time of the server farm for $n_1 = 2, 3, 4, 5$.

(c) Based on your answer to Parts (a) and (b), determine the value of n_1 that gives the minimum response time for the server farm.

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 $---$