COMP9334 Capacity Planning
Assignment, Session 1, 2015

March 26, 2015

Instructions

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

(2) The total marks for this assignment is 15 marks.

(3) In answering the questions, it is important for you to show us your intermediate steps and tell us 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, but 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 are also required to submit the program.

(4) The submission deadline is 11:59pm Sunday, 19 April 2015 (i.e. the day before Monday of Week 7). Late submission will cap the maximum mark that you receive.

(5) Submit your solution via *give* command. We will only accept Acrobat pdf file with the name *assign.pdf*. Log onto a CSE machine and make sure you're in the same directory as your work, then do the following:

(a) When you're ready to submit, at the bash prompt type 9334

(b) Next, type: give cs9334 assign assign.pdf (You should receive a message stating the result of your submission).

Note that 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 three devices: a CPU and three disks (denoted by Disk1, Disk2 and Disk3). The system was monitored for 30 minutes and the following measurements were taken:

Number of completed jobs	1,650
Number of CPU accesses	3,000
Number of Disk1 accesses	19,500
Number of Disk2 accesses	13,200
Number of Disk3 accesses	25,200
CPU busy time	1213 seconds
Disk1 busy time	1027 seconds
Disk2 busy time	937 seconds
Disk3 busy time	1356 seconds
Think time	31 seconds

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

(b) Use bottleneck analysis to determine the asymptotic bound on the system throughput when there are 50 active terminals.

(c) Using your results in Part (b), compute the minimum possible response time when the number of terminals is 50.

Question 2 (6 marks)

Consider a server farm shown in Figure 1. The server farm consists of n CPUs (where n is a positive integer greater than 1) and a dispatcher.

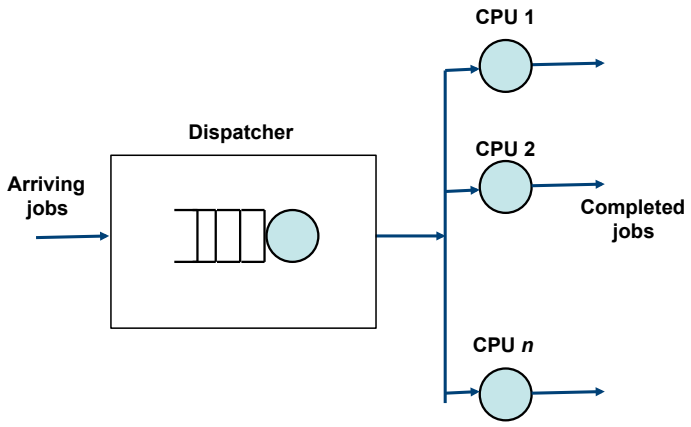

Figure 1: Figure for Question 2.

Jobs arrive at this server farm with an inter-arrival time which is exponentially distributed with a mean arrival rate 17 jobs per second. The processing time that each job requires at a CPU is exponentially distributed and a CPU can process on average of 9 jobs per second. There are no queues at the CPUs.

The dispatcher has a queue which can store up to 10 jobs. When a job arrives at the server farm, the dispatcher executes the following algorithm. Note that the statement behind % is a comment.

if the queue (at the dispatcher) is full
    Reject the job
else % The queue is not full and the job will be accepted
    if the dispatcher queue is empty
        if all CUPs are busy
            The job joins the end of the queue at the dispatcher
        else % At least one CPU is idle
            The job is sent to any idling CPU for processing
    else % The dispatcher queue is not empty
        The job joins the end of the queue at the dispatcher

When a CPU has finished processing a job, it will take the job from the head of the queue (if there is one) at the dispatcher and process it.

You may assume that the dispatcher takes a negligible amount of time to check whether a CPU is available. You may also assume that it takes negligible time for a CPU to check whether there are jobs in the dispatcher and to transfer it to the CPU for processing. These assumptions mean that the processing time in the server farm is dominated by the processing time at the CPU.

Answer the following questions:

(a) The computation system described above can be modelled by a queueing system. How will you describe this queueing system in Kendall's notation?

(b) Formulate a continuous-time Markov chain for the system described above. Your formulation should include the definition of the states and the transition rates between states.

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

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

(e) Assuming the current system consists of 2 CPUs, determine:

  (i) The probability that an arriving job will not be dropped (i.e. will be accepted).

  (ii) The mean waiting time of an accepted job in the queue. Let us denote the result of this by x.

(f) Assuming that you are the administrator of the computation system and you are not satisfied with the current system (which consists of 2 CPUs) because the waiting time of an accepted job is too high. You decide to add more CPUs to the system so that the waiting time becomes less than $0.1x$. Determine the minimum number of CPUs that you need to add.

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 (6 marks)

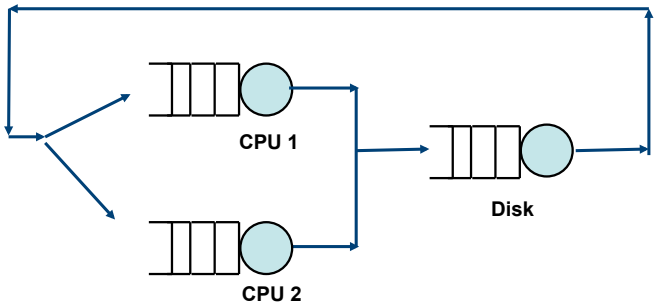

Figure 2: Figure for Question 3.

Consider the computer system shown in Figure 2. The system consists of three devices: a disk and 2 CPUs. Each device is modelled as a server and a queue. The system is at peak load and there are four (4) jobs circulating in the system at all times. During each round that a job circulates the system, the job requires processing from *one* of the CPUs and then followed by the disk. Assuming that:

- The processing time required by each job per visit to the disk is exponentially distributed with mean 100 milli-seconds.
- The two CPUs have different mean processing times. The mean processing times for CPU1 and CPU2 are, respectively, 100 and 200 milli-seconds. Both processing time distributions are assumed to be exponential.
- After a job has left the disk, it will proceed to receive processing at one of the CPUs immediately. In any attempt to utilise the faster CPU (i.e. CPU1), the following job assignment strategy is employed:
  (a) If both CPU1 and CPU2 are idle, the job will be sent to CPU1.
  (b) If only CPU2 is idle, the job will be sent to CPU2.
  (c) If CPU2 is busy, the job will be sent to CPU1.

Answer the following questions.

(a) Let the states be the following 3-tuple:

(number of users in the CPU1, number of users in CPU2, number of users in the disk),

formulate a continuous-time Markov chain for this computer system. Your formulation should include (1) a list of states; (2) the transition rates between the states.

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

(c) What are the steady state probabilities for each state?

(d) What is the throughput of the system?

(e) What is the mean response time of the CPU1?

(f) How long does a user have to wait, on average, at the disk before it gets served?

Note: If you use a computer program to solve for the steady state probabilities, you need to show us your code. Also, do not forget to show us the steps you use to get your answers.

$---$ End of assignment $---$