Homework #4 (100 Points)

Due 03:00pm on Tuesday 03/01/2016 through gsubmit.

Feel free to make assumptions, if you feel that such assumptions are justified or necessary. Please state your assumptions clearly. Unreasonable assumptions lead to unreasonable grades!

---

1. The buffer of a streaming media application can hold up to K packets. The processing time for each packet to be played out was found to be exponential with a mean of 20 msec. Degradation in the quality of the playout occurs as a result of buffer over-runs or as a result of a buffer under-runs. Buffer over-runs occur when a packet is dropped due to a filled buffer (causing a "blip" in the playout). Buffer under-runs occur when there are no packets to playout (causing a period of "plop" in the playout). It was measured that the packets arrive to the application as a Poisson process with a mean of 50 packets per second. Answer the following questions:

    a. Explain why it makes sense to model the above system as an M/M/1/K queue.
    b. Assuming K=3 packets, how many packets do you expect to find in the buffer (i.e. waiting to be played out)?
    c. Assuming K=3 packets and that a streaming media object consists of 2,000 packets, how many "blips" and how many "plops" do you expect to hear for that object?
    d. Assuming K=3 packets, what is the mean delay between receipt of a packet and completing its playout?
    e. Assuming K=10 packets, what is the mean delay between receipt of a packet and completing its playout?
    f. Compare your answers for part (d) and part (e). Provide an explanation for the difference. In general, do you expect the mean delay for packets that are played out to increase or decrease as a function of K?
    g. Based on your answers above, provide a justification for the following statement: "Increasing the size of network buffers (i.e., K) has its advantages and its disadvantages".
    h. Would you recommend a large or a small value of K for a "Movie on Demand" streaming application (e.g., Netflix)? Will your answer be different for a "Teleconferencing" streaming application (e.g., Skype)? Justify your answers.

2. Modify the M/M/1 discrete event simulator that you that you developed for Assignment #3 to simulate an M/M/1/K queue (i.e., a queue with a finite buffer, where q <= K). Clearly, the difference between an M/M/1/K queue and an M/M/1 queue is that upon a "birth event", one of two things may happen, either the queue is not full (i.e., total number of requests in the system q is less than K), in which case the birth event results in the addition of one more customer to the queue (as in the M/M/1 queue), or else, the queue is full, in which case the birth event is "rejected".

    To test your simulator, use it to verify the following settings for an M/M/1/K system

    a. K = 3, Lambda = 50, Ts = 0.020 and simulation time = 100

b. K = 3, Lambda = 50, Ts = 0.015 and simulation time = 100
c. K = 3, Lambda = 65, Ts = 0.015 and simulation time = 100

For each of the above report the results you obtain from simulation for the rejection probability and for the 95th-percentile confidence interval for q and Tq. Compare these results to what you expect to get from analysis.

Modify your M/M/1/K simulator so as to simulate an M/D/1/K system -- i.e., a finite queue of size K with Poisson arrivals and a fixed service time. To test your simulator, use it to verify the following settings for an M/D/1/K system:

d. K = 3, Lambda = 50, Ts is fixed and equal to 0.015, and simulation time = 100
e. K = 3, Lambda = 65, Ts is fixed and equal to 0.015, and simulation time = 100

For each of the above report the results you obtain from simulation for the rejection probability and for the 95th-percentile confidence interval for q and Tq. Compare the results you obtain to those you got for parts (a) and (b). In particular, is the rejection probability more or less? Is that what you expected? Explain.

Note: The setting for part (a) in this problem is the same as the setting used in problem #1 above. The settings for parts (b) and (c) in this problem are comparable to the settings in parts (b) and (c) for the M/M/1 system you simulated in Assignment #3. You may want to compare the results in both cases to make sure they make sense. Do the results make sense?

What to hand in: Source code that we can compile and test; answers to parts a-e showing your work.

3. Consider the system described in problem #4 in Assignment #3, but with the following assumptions:
   - Process arrivals are Poisson with a rate of 40 processes per second
   - CPU service time is uniformly distributed between 1 and 39 msec
   - Disk I/O service time is normally distributed with mean of 100 msec and standard deviation of 30 msec (but never negative)
   - Network service time is constant and equal to 25 msec
   - All buffers are of infinite size

   Write a discrete-event simulator for the system described above. Run your simulator for a period of 100 seconds and answer the following:

   a. Plot the number of requests waiting in the CPU queue as a function of time.
      Note: Obtain a measurement using PASTA principle every (simulated) second on average and plot your measurement as a function of time (i.e., from time 0 to time 100) using spreadsheet software such as Excel.
   b. Determine a point in the simulation after which you feel confident that the system has reached a steady state.

c. Compute the 95th and 98th percentile confidence interval for the:
   - Number of requests in the CPU queue at steady state
   - Response time through the entire system at steady state

d. How do the results you obtain from this simulation compare to those you got analytically in problem #4 in Assignment #3? Comment on whether any similarities or differences make sense.

What to hand in: Source code that we can compile and test; answers to parts a-d showing your work.

4. Consider the same system used for the above problem, answer the following questions:
   a. Estimate the slowdown of the system. Recall that the "slowdown" for any system is defined as the ratio of the mean (actual) response time (at steady state) to the minimum response time. One way to figure out the minimum response time is to simulate the system with a very low arrival rate (e.g., one request per second as opposed to 40 requests per second).
   b. Estimate the capacity of the system. Recall that the "capacity" of any system is defined as the maximum rate at which the system is able to serve requests submitted to it. One way to figure out the capacity of the system via simulation is to simulate the system repeatedly, while increasing the arrival rate in successive simulations until any one of the resources in the system reaches its breaking point, i.e., utilization close to 100%.

What to hand in: Source code that we can compile and answers to parts a and b showing your work.