CS5239 Assignment

The assignment must be submitted on Canvas as answers to quiz Assignment by 30 Sep, 6:30pm

You have one attempt and 60 minutes to complete the quiz. Questions that need computation and/or explanation are shown below. The quiz contains 10 other multiple-choice questions not shown in this file (numbered questions 1 to 10 in the quiz), adding up to [10 marks]. The assignment covers lectures 1-5. The Canvas quiz might show the options for the MCQ questions in a different order compared to what we see in this document.

Two sets of practice questions have been published in Canvas Files. Practice your problem-solving skills with these sets before attempting the assignment.

To have enough time to answer the questions in the quiz, answer and type your answers to the questions below before starting your quiz attempt. Note that 50 minutes is not enough to read, solve, and type your answers to all the questions in the quiz. It would be best if you prepared your answers before attempting the quiz.

You need to show your work in the Canvas Quiz only for the questions that specifically say "Justify...", or "Explain ..." or "Show your work." in the PDF. You may type out the explanation or upload an image with the derivation. For the other questions, you need to give only the final answer.

Question 11. [2 marks] In a computer system, there are on average 30 users in the queue. If the arrival rate of users is 16 users per hour and on average 32 users leave the system per hour, what is the average number of users in the system?

a. 32
b. 16.5
c. 30.5
d. 46

Question 12. [2 marks] In a single-server datacenter with mean arrival rate of λ and mean service time of $1/\mu$, the expected number of customers in the datacenter is $\lambda/(\mu - \lambda)$. What is the expected response time per user in the datacenter?

a. $\lambda^2/(\mu - \lambda)$
b. $\mu - \lambda$
c. $1/(\mu - \lambda)$
d. $(\mu - \lambda)/\lambda$

Question 13. [3 marks] For a small web-server with just one processor, user transactions arrive at a rate of 1,500 transactions per hour. The web-server executes user transactions with a mean service time of 1 second on average. What is the mean response time of user transactions?

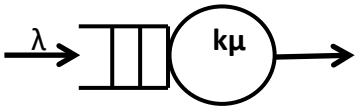
Question 14. [2 marks] A system programmer observes a batch processing job queue in a computer system for a period of one day. On average there are 100 jobs in the queue at any one time and 50 job completes execution per second. What is the average time it takes a job to complete processing?

Question 15. [2 marks] Consider an M/M/1 queueing system. Customer interarrival times have an average of 5 minutes, and service times have an average of 4 minutes. What will be the average number of customers waiting in line? (Give your answer including one decimal place)

Question 16. [1 mark] Consider an M/M/3 queueing system having 10 arrivals per hour and a mean service time of 10 minutes. What is the traffic intensity of this system? (Give your answer including two decimal places)

Answer questions 17-24 using the following problem description: (Spacer in quiz)

A system manager currently operates a single-core mainframe computer system with job arrival rate λ and a service rate of the single-core is kμ. This single-core mainframe computer system can be modeled as an M/M/1 system shown below.



Two alternative configurations are proposed by vendors A and B and the system manager would like to evaluate these configurations using *mean response time*.

- Vendor A (distributed servers) proposes a system with *k smaller servers* with arrival and departure rates divided into k streams of λ/k and μ respectively. In this configuration, each server has its own queue.

Vendor B (multi-core processor) proposes a configuration with *k smaller servers.* All job arrivals join a *single queue* that serves *k servers* with a service rate of μ.

Question 17. [1 mark] What is a suitable queueing model for vendor A?
Questions 18 & 19. [2 marks] Using mean response time, compare the original M/M/1 system with vendor A system. Based on this comparison, which system has a lower performance? Show your work and justify your answer.
Question 20. [1 mark] What is a suitable queueing model for vendor B?
Questions 21 & 22. [2 mark] Using mean response time, how does the original M/M/1 system compare with vendor B system when the *workload is low*? Show your work and justify your answer.
Questions 23 & 24. [2 mark] Using mean response time, how does the original M/M/1 system compare with vendor B system when the *workload is high*? Show your work and justify your answer.

Answer questions 25-29 using the following problem description: (Spacer in quiz)

EasyTest, a health diagnostic SME, runs a web server consisting of a cluster of servers and an array of disks. As the company's performance analyst, you are tasked by management to evaluate the computing capacity of the company's web server.

You choose to use operational analysis to evaluate the performance of this web server. For simplicity, you model (approximate) the cluster of servers and the disk array as a large single **CPU** server and as a large single **disk** respectively. Since http requests have different processing demands, these requests (workload) are divided into two types: html file requests and image file requests.

To obtain the server workload, you observed (measured) the request arrivals at the web server for **three hours.** In this observation period, there are 42,000 html and 3,100 image file requests and the size of html files are 3,000 bytes and image files are 15,000 bytes on average.

From computer networking, CPU service demand for an http request is divided into two parts: *constant* overhead for processing a request (includes open TCP connection, analyze the http request and open the requested file) and processing the file size of each request (CPU is involved in each I/O operation). For simplicity, we model CPU demand, in seconds, per http requests, as CPUDemand = 0.008 + (0.002 x request_size) where 0.008 seconds is a constant time overhead, 0.002 is the service time per block read, request_size is the number of blocks read and each block read is 1,000 bytes. The average disk demand is 12 msec for each 1,000-byte block.

Question 25. [2 marks] Let's model this system as an *open* queueing network. What are the *average arrival rates* (in requests/second) for each type of request – λ_{html} and λ_{image}?

Question 26. [4 marks] For each type of request, compute the **service demand** (in seconds) and **utilization** at the CPU and disk. For example, service demand, Dhtml, cpu, Dhtml, disk, Dimage, cpu, and Dimage, disk. (3 decimal places)

Question 27. [1 mark] What is the bottleneck workload?

Question 28. [1 marks] What is the bottleneck device, i.e., CPU or disk?

Question 29. [2 marks] Compute the average response time (in seconds) for a html and an image request (3 decimal places)