Depth Screening Exam
Modeling and Analysis of Computer Systems

University of Wisconsin
February 1990

Directions: Answer four of the first five questions, and questions 6-8. Define the notation you use, the assumptions you make, and/or explain your reasoning in deriving each answer whenever appropriate.

Question 1

Consider a *closed* queueing network consisting of just one single-server FCFS center. Service times are independent and identically distributed with mean E(X) and distribution function $F_X(x)$. What is the throughput and mean response time as a function of the customer population, N?

Question 2

(a) Consider a queueing network with three queues in tandem (*i.e.*, three queues connected serially). Suppose that the service rates of the three servers, μ_1, μ_2, and μ_3, are constrained such that $\mu_1 + \mu_2 + \mu_3 = \mu$. For what values of the three service rates (in terms of μ) is the system response time minimized? Explain why the response time is minimized at these values.

(b) For the queueing network in part (a), will minimizing the response time have greater impact if the queueing network is open or if it is closed? Explain.

Question 3

Give the Mean Value Analysis equations for mean response time, throughput, and mean queue length at a FCFS queue in an open single-class queueing network. Give the corresponding equations for a PS queue in the same network. Simplify the equations for mean queue length, to get an expression in terms of the input parameters of the model.

Question 4

Consider the single-class m-server queue, in which the rate of each server is $\frac{\mu}{m}$. If the arrival process is Poisson and service demand is exponentially distributed, then mean response time is minimized at $m = 1$. For what workload characteristics might mean response time be minimized at some value $m > 1$? Under these circumstances, suggest a reasonable hypothesis for how mean response time might vary, qualitatively, with m. Explain the reasoning behind your hypothesis.

Question 5

(a) What is the difference between a *discrete time* model and a *discrete event* model?

(b) Can two events take place at the same simulation time in a discrete event model? If the answer is **no** explain why, if the answer is **yes**, explain how the event scheduler of a simulation system should decide which event should be executed first.

(c) When does a subset of the descriptive variables of a discrete event model constitute a set of state variables for the model?

(d) What is the difference between an *incomplete* model and an *inconsistent* model?

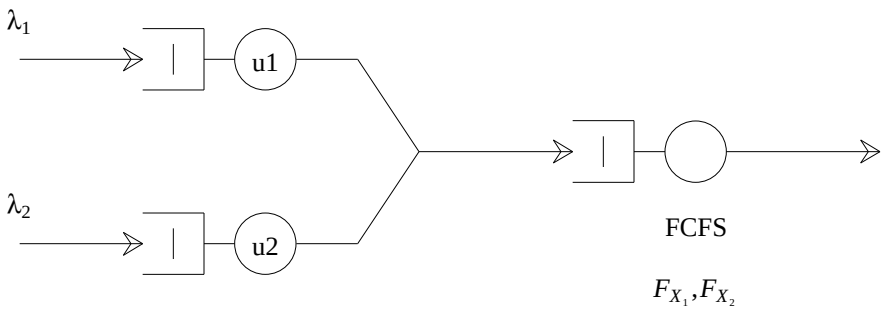
Question 6

Consider a single-class *m*-server queue, all servers equally fast. Service times for the jobs are independent and identically distributed. Let *X* denote the service time of a job. Each job consists of exactly *m* independent tasks that can be executed in parallel, each of which has service requirement equal to $\frac{X}{m}$. Assuming the scheduling discipline for the queue is FCFS, derive the mean response time of the queue. (Hint: Think carefully about the behavior of the queue.)

Question 7

Consider the open queueing network model below. Class 1 arrives to queue 1 at rate λ_1 and class 2 arrives to queue 2 at rate λ_2. Each arrival process is a Poisson process. Service times at queue 1 are independent and exponentially distributed with rate μ_1; service times at queue 2 are independent and exponentially distributed with rate μ_2. A departure from either queue 1 or queue 2 proceeds to queue 3. The scheduling discipline at queue 3 is FCFS. Each class has its own general service time distribution $F_{X_i}(x)$ with mean $\overline{X_i}$, $i = 1,2$.

Derive an expression for the mean waiting time at queue 3. Is your expression exact or approximate? Explain.



Question 8

Assume a trace driven simulation model of a single-server queue. The scheduling discipline of the queue is a parameter of the model. The only output of the model is the expected response time of the system, R.

The system was simulated three times with the same trace and with three different scheduling disciplines. Let W_{FIFO}, W_{PS}, and W_{random} be the observed mean response times for the first in first out, processor sharing, and random disciplines respectively. The results for the trace simulations were $W_{FIFO} < W_{random} < W_{PS}$. Explain *carefully* what these results indicate about the characteristics of the traced workload. Describe a measurement procedure that will verify your hypotheses.