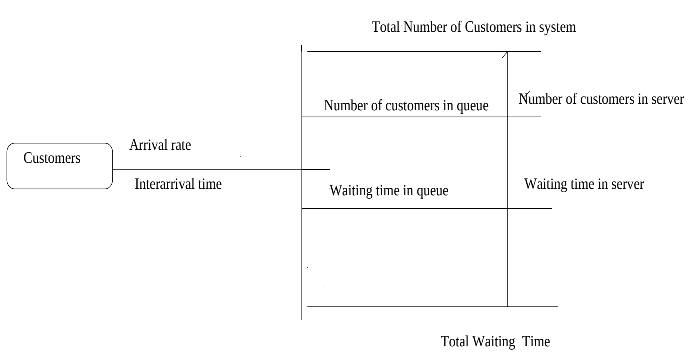
Activity 2. Theoretical Problem

Queueing theory is a very important modeling technique that represents a computer system as a network of service centers, each of which is treated as a queueing system. That is, each service center has an associated queue or waiting line where customers who cannot be served immediately queue (wait) for service. The customers are, of course, part of the queueing network. Customer is a generic word used to describe workload requests such as CPU service, I/O service requests, requests for main memory, etc. All these arrive at random to the service facility. Queueing theory models are often used to determine the effects of changes in the configuration of a computer system.

The simplest queueing theory model is the `M/M/1` model, which assumes: (a) that customers arrive in accordance with a Poisson process with average rate λ and thus the inter-arrival times are exponentially distributed with mean $\frac{1}{\lambda}$. (b) Service time by the server is assumed exponential with parameter μ. Expected service time is then $\frac{1}{\mu}$. (c) The customers are served one at a time by a single server. If the server is busy upon the customer's arrival, then the customer waits in the queue. The activity presented below assumes that the random variables that comprise this model were used as examples when talking about random variables in class, and when talking about the exponential and the Poisson random variables. The set of random variables involved is summarized in the diagram displayed below, which is an adaptation of a diagram in Allen(1990), p. 251.

Queueing Theory random variables. A queueing system.



For these models we are usually interested in determining, among other things, the average number of customers in the system (or in the queue) and the average amount of time a customer spends in the system.

From the assumptions above, it can be proved (Allen 1990) that the number of customers in the system has a geometric distribution with parameter $p = 1 - \frac{\lambda}{\mu}$ and $q = \frac{\lambda}{\mu}$. So the expected number of customers is $\frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}}$. And the probability

$$\begin{aligned} P(N < n) &= P(N = 0) + \cdots + P(N = n-1) \\ &= 1 - (\frac{\lambda}{\mu})^n \end{aligned}$$

Consequently,

$$P(N \geq n) = 1 - P(N < n) = (\frac{\lambda}{\mu})^n.$$

The following problem, inspired by Allen(1990) p. 267, makes use of this result. See Appendix B for a brief discussion of its solution. More detailed solutions will appear in the final Solutions Manual.

Problem

Traffic to an e-mail server arrives in a random pattern (i,e, exponential interarrival time) at a rate of 240 e-mails per minute. The server has a transmission rate of 800 characters per second. The message length distribution (including control characters) is approximately exponential with an average length of 176 characters. Assume a `M/M/1` queueing system like that often used in class when talking about random variables (i.e., exponential arrival times, exponential service time, and 1 server). What is the probability that 10 or more messages are waiting to be transmitted? Support your answer showing work and explaining the implications of the assumptions.

References

Allen, A.O.(1990). Probability, Statistics, and Queueing Theory with Computer Science Applications, 2nd edition. Academic Press.