Queuing Theory

Stochastic Processes

Fatih Cavdur
fatihcavdur@uludag.edu.tr

November 7, 2015

Introduction

- We study a class of models in which customers arrive in some random manner at a service facility.
- Upon arrival they are made to wait in queue until it is their turn to be served.
- Once served, they are generally assumed to leave the system.
- For such models, we are interested in determining, among other things, such quantities as the average number of customers in the system (or in the queue) and the average time a customer spends in the system (or in the queue).

Preliminaries

Some fundamental quantities of interest are

- L: the average number of customers in the system
- L_Q: the average number of customer in the queue
- W: the average amount of time a customer spends in the system
- W_Q: the average amount of time a customer spends in the queue

Introduction

Basic cost identity can be stated as

$$\begin{pmatrix} \text{average} \\ \text{rate} \\ \text{at which} \\ \text{system} \\ \text{earns} \end{pmatrix} = \begin{pmatrix} \text{avarage} \\ \text{arrival} \\ \text{rate of} \\ \text{entering} \\ \text{customers} \end{pmatrix} \times \begin{pmatrix} \text{average} \\ \text{amount} \\ \text{an entering} \\ \text{customer} \\ \text{pays} \end{pmatrix}$$

where average arrival rate of entering customers (λ_a) is defined as

$$\lambda_a = \lim_{t \to \infty} \frac{N(t)}{t}$$

Introduction

Little's Formula

$$L = \lambda_a W$$

$$L_Q = \lambda_a W_Q$$

Introduction

By assuming a unit cost per unit time (i.e., \$1 per unit time) while in service, we can write from the basic cost identity that

$$\begin{pmatrix} \text{average number of customers} \\ \text{in service} \end{pmatrix} = \lambda_a E(S)$$

where $E(S)$ is defined as the average amount of time a customer spends in service.

Steady-State Probabilities

Let $X(t)$ be the number of customers in the system at time t, and define P_n, $n \geq 0$ by

$$P_n = \lim_{t \to \infty} P\{X(t) = n\}$$

where we assume the limit exists. We can say that P_n is the limiting or long-run or steady-state probability that there will be exactly n customers in the system. We also define the other limiting probabilities $\{a_n, n \geq 0\}$ and $\{d_n, n \geq 0\}$ where

- a_n: proportion of customers that find n in the system when they arrive
- b_n: proportion of customers that leave behind n in the system when they depart

A Single-Server Exponential Queuing System

For such a queuing system, we can write

state	rate the process leaves	rate the process enters
0	λP_0	μP_1
$n, n \geq 1$	$(\lambda + \mu) P_n$	$\lambda P_{n-1} + \mu P_{n+1}$

In order to solve the *balance equations*, we can write

$$P_1 = \frac{\lambda}{\mu} P_0$$

$$P_{n+1} = \frac{\lambda}{\mu} P_n + \left(P_n - \frac{\lambda}{\mu} P_{n-1} \right), \quad n \geq 1$$

A Single-Server Exponential Queuing System

Solving in terms of P_0, we get

$$P_1 = \frac{\lambda}{\mu} P_0$$

$$P_2 = \frac{\lambda}{\mu} P_1 + \left(P_1 - \frac{\lambda}{\mu} P_0 \right) = \frac{\lambda}{\mu} P_1 = \left(\frac{\lambda}{\mu} \right)^2 P_0$$

$$P_3 = \frac{\lambda}{\mu} P_2 + \left(P_2 - \frac{\lambda}{\mu} P_1 \right) = \frac{\lambda}{\mu} P_2 = \left(\frac{\lambda}{\mu} \right)^3 P_0$$

$$P_{n+1} = \frac{\lambda}{\mu} P_n + \left(P_n - \frac{\lambda}{\mu} P_{n-1} \right) = \frac{\lambda}{\mu} P_n = \left(\frac{\lambda}{\mu} \right)^{n+1} P_0$$

A Single-Server Exponential Queuing System

To determine P_0,

$$\sum_{n=0}^{\infty} P_n = \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n P_0 = \frac{P_0}{1-\lambda/\mu} = 1 \Rightarrow P_0 = 1 - \frac{\lambda}{\mu}$$

$$P_0 = 1 - \frac{\lambda}{\mu} \Rightarrow P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right), \; n \geq 1$$

and L,

$$L = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) = \frac{\lambda}{\mu} \Rightarrow W = \frac{L}{\lambda} = \frac{1}{\mu - \lambda}$$

A Single-Server Exponential Queuing System

Finally,

$$W_Q = W - E(S) = W - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu - \lambda)}$$

and

$$L_Q = \lambda W_Q = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

A Single-Server Exponential Queuing System with Finite Capacity

For such a queuing system, we can write

state	rate the process leaves	rate the process enters
0	λP_0	μP_1
$1 \leq n \leq N-1$	$(\lambda+\mu)P_n$	$\lambda P_{n-1} + \mu P_{n+1}$
N	λP_{N-1}	μP_N

In order to solve the *balance equations*, we can write

$$P_n = \frac{\lambda}{\mu} P_{n-1} = \left(\frac{\lambda}{\mu}\right)^2 P_{n-2} = \left(\frac{\lambda}{\mu}\right)^n P_0, \;\; n = 1, \ldots, N$$

A Single-Server Exponential Queuing System with Finite Capacity

To determine P_0,

$$\sum_{n=0}^{\infty} P_n = \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n P_0 = P_0 \left[\frac{1-(\lambda/\mu)^{N+1}}{1-\lambda/\mu}\right] = 1$$

We then have,

$$P_0 = \frac{1-\lambda/\mu}{1-(\lambda/\mu)^{N+1}} \Rightarrow P_n = \frac{(\lambda/\mu)^n(1-\lambda/\mu)}{1-(\lambda/\mu)^{N+1}}, \;\; n = 0, \ldots, N$$

A Single-Server Exponential Queuing System with Finite Capacity

To determine L,

$$
\begin{aligned}
L &= \sum_{n=0}^{\infty} nP_n \\
&= \frac{(1-\lambda/\mu)}{1-(\lambda/\mu)^{N+1}} \sum_{n=0}^{\infty} n \left(\frac{\lambda}{\mu}\right)^n \\
&= \frac{\lambda[1+N(\lambda/\mu)^{N+1}-(N+1)(\lambda/\mu)^N]}{(\mu-\lambda)[1-(\lambda/\mu)^{N+1}]}
\end{aligned}
$$

A Shoeshine Shop

Consider a shoeshine shop consisting of two chairs. Suppose that an entering customer first will go to chair 1. When his work is completed in chair 1, he will go either to chair 2 if that chair is empty or else wait in chair 1 until chair 2 becomes empty. Suppose that a potential customer will enter this shop as long as chair 1 is empty. (Thus, for instance, a potential customer might enter even if there is a customer in chair 2.)
If we suppose that potential customers arrive in accordance with a Poisson process at rate λ , and that the service times for the two chairs are independent and have respective exponential rates of μ_1 and μ_2, then

- what proportion of potential customers enters the system?
- what is the mean number of customers in the system?
- what is the average amount of time that an entering customer spends in the system?

A Shoeshine Shop

To begin we must first decide upon an appropriate state space. It is clear that the state of the system must include more information than merely the number of customers in the system. For instance, it would not be enough to specify that there is one customer in the system as we would also have to know which chair he was in. Further, if we only know that there are two customers in the system, then we would not know if the man in chair 1 is still being served or if he is just waiting for the person in chair 2 to finish.

A Shoeshine Shop

We have the following state definitions:

- state $(0,0)$: no customers in the system
- state $(1,0)$: 1 customer in chair 1
- state $(0,1)$: 1 customer in chair 2
- state $(1,1)$: 2 customers both in service
- state $(b,1)$: 2 customers, first is done, second in service

A Shoeshine Shop

We can then write

$$\begin{aligned}
(0,0) &\Rightarrow \lambda P_{00} = \mu_2 P_{01} \\
(1,0) &\Rightarrow \mu_1 P_{10} = \lambda P_{00} + \mu_2 P_{11} \\
(0,1) &\Rightarrow (\lambda + \mu_2) P_{01} = \mu_1 P_{10} + \mu_2 P_{b1} \\
(1,1) &\Rightarrow (\mu_1 + \mu_2) P_{11} = \lambda P_{01} \\
(b,1) &\Rightarrow \mu_2 P_{b1} = \mu_1 P_{11}
\end{aligned}$$

and

$$P_{00} + P_{10} + P_{01} + P_{11} + P_{b1} = 1$$

A Shoeshine Shop

After solving the equations, we can write

$$L = P_{10} + P_{01} + 2(P_{11} + P_{b1})$$

and

$$W = \frac{P_{10} + P_{01} + 2(P_{11} + P_{b1})}{\lambda(P_{00} + P_{01})}$$

Open Systems

We have, for a 2-server open system,

$$\begin{aligned}
(0,0) &\Rightarrow \lambda P_{00} = \mu_2 P_{01} \\
(n,0) &\Rightarrow (\lambda + \mu_1) P_{n0} = \mu_2 P_{n1} + \lambda P_{n-1,0} \\
(0,m) &\Rightarrow (\lambda + \mu_2) P_{0m} = \mu_2 P_{0,m+1} + \mu_1 P_{1,m-1} \\
(n,m) &\Rightarrow (\lambda + \mu_1 + \mu_2) P_{nm} = \mu_2 P_{n,m+1} + \mu_1 P_{n+1,m-1} + \lambda P_{n-1,m}
\end{aligned}$$

and

$$\sum_n \sum_m P_{nm} = 1$$

Open Systems

We first note that, we have an $M/M/1$ model for the first server, and since, the departure is a Poisson process with rate λ, we have an $M/M/1$ model for the second server also. We thus have,

$$P\{N_1 = n\} = \left(\frac{\lambda}{\mu_1}\right)^n \left(1 - \frac{\lambda}{\mu_1}\right)$$

and

$$P\{N_2 = m\} = \left(\frac{\lambda}{\mu_2}\right)^m \left(1 - \frac{\lambda}{\mu_2}\right)$$

Open Systems

If the numbers of customers in the first and second servers are independent random variables,

$$P_{nm} = \left(\frac{\lambda}{\mu_1}\right)^n \left(1 - \frac{\lambda}{\mu_1}\right) \left(\frac{\lambda}{\mu_2}\right)^m \left(1 - \frac{\lambda}{\mu_2}\right)$$

By solving the balance equations, we can show, for instance,

$$\lambda \left(1 - \frac{\lambda}{\mu_1}\right) \left(1 - \frac{\lambda}{\mu_2}\right) = \mu_2 \left(1 - \frac{\lambda}{\mu_1}\right) \left(\frac{\lambda}{\mu_2}\right) \left(1 - \frac{\lambda}{\mu_2}\right)$$

Open Systems

We can then write,

$$\begin{aligned} L &= \sum_n \sum_m (n+m) P_{nm} \\ &= \sum_n n \left(\frac{\lambda}{\mu_1}\right)^n \left(1 - \frac{\lambda}{\mu_1}\right) + \sum_m m \left(\frac{\lambda}{\mu_2}\right)^m \left(1 - \frac{\lambda}{\mu_2}\right) \\ &= \frac{\lambda}{\mu_1 - \lambda} + \frac{\lambda}{\mu_2 - \lambda} \Rightarrow W = \frac{L}{\lambda} = \frac{1}{\mu_1 - \lambda} - \frac{1}{\mu_2 - \lambda} \end{aligned}$$

Open Systems

By generalizing the preceding, we have

$$\lambda_j = r_j + \sum_i^k \lambda_i P_{ij},\ \forall j \Rightarrow P\{N_j = n\} = \left(\frac{\lambda_j}{\mu_j}\right)^n \left(1 - \frac{\lambda_j}{\mu_j}\right),\ n \geq 1$$

and

$$P(n_1, n_2, \ldots, n_k) = \prod_{j=1}^k \left(\frac{\lambda_j}{\mu_j}\right)^{n_j} \left(1 - \frac{\lambda_j}{\mu_j}\right)$$

and

$$L = \sum_{j=1}^k \frac{\lambda_j}{\mu_j - \lambda_j} \Rightarrow W = \frac{\sum_{j=1}^k \lambda_j/(\mu_j - \lambda_j)}{\sum_{j=1}^k r_j}$$

Closed Systems

We assume that we have m customers moving among a system of k servers where the service at server i is exponential with rate μ_i, $i = 1, \ldots, k$. When a customer completes service at service i, she then joins the queue in front of server j, $j = 1, \ldots, k$, with probability P_{ij}. We assume that $\mathbf{P} = [P_{ij}]$ is a Markov TPM. We also assume that it is irreducible.

Closed Systems

We can then show that

$$P_m(n_1, \ldots, n_k) = \begin{cases} C_m \prod_{j=1}^{k} (\pi_j/\mu_j)^{n_j}, & \text{if } \sum_{j=1}^{k} n_j = m \\ 0, & \text{otherwise} \end{cases}$$

where

$$C_m = \left(\sum_{\substack{n_1, \ldots n_k \\ \sum_j n_j = m}} \left[\prod_{j=1}^{k} \left(\frac{\pi_j}{\mu_j} \right)^{n_j} \right] \right)^{-1}$$

General Service Time Distribution

For an arbitrary queuing system, let X be the amount paid by a customer, and we can then write

$$V = \lambda_a E(X)$$

and then

$$\begin{aligned} E(X) = E\left[SW_Q^* + \int_0^S (S - x)dx \right] \Rightarrow V &= \lambda_a E(SW_Q^*) + \frac{\lambda_a E(S^2)}{2} \\ &= \lambda_a E(S) W_Q + \frac{\lambda_a E(S^2)}{2} \end{aligned}$$

General Service Time Distribution

For an $M/G/1$ system, since we have

$$W_Q = V \quad \text{and} \quad V = \lambda E(S) W_Q + \frac{\lambda E(S^2)}{2}$$

we can write the following Pollackzek-Khintchine formula as follows.

$$W_Q = \frac{\lambda E(S^2)}{2[1 - \lambda E(S)]}$$

General Service Time Distribution

Using the following Pollackzek-Khintchine formula

$$W_Q = \frac{\lambda E(S^2)}{2[1 - \lambda E(S)]}$$

we can write

$$L_Q = \lambda W_Q = \frac{\lambda^2 E(S^2)}{2[1 - \lambda E(S)]}$$

$$W = W_Q + E(S) = \frac{\lambda E(S^2)}{2[1 - \lambda E(S)]} + E(S)$$

$$L = \lambda W = \frac{\lambda^2 E(S^2)}{2[1 - \lambda E(S)]} + \lambda E(S)$$

General Time-Between Arrivals Distribution

For an $G/M/1$ system, we have

$$W = \frac{1}{\mu(1-\beta)}$$

$$W_Q = \frac{\beta}{\mu(1-\beta)}$$

$$L = \frac{\lambda}{\mu(1-\beta)}$$

$$L_Q = \frac{\lambda\beta}{\mu(1-\beta)}$$

Machine Repair Model

We can show, for the machine repair model, that

$$L_Q = \frac{m(m-1)\mu_R - m(1-\pi_0)/\lambda}{m\mu_R + \pi_0/\lambda}$$

and

$$L = \frac{m^2\mu_R - m(1-\pi_0)/\lambda}{m\mu_R + \pi_0/\lambda}$$

Multi-Server Queues

- $M/M/k$
- $G/M/k$
- $M/G/k$

The End

fatihcavdur@uludag.edu.tr