Chapter 4
Queueing Theory

The priest persuades humble people to endure their hard lot, a politician urges them to rebel against it, and a scientist thinks of a method that does away with the hard lot altogether.

—Max Percy

Queueing is simply waiting in lines such as stopping at the toll booth, waiting in line for a bank cashier, stopping at a traffic light, waiting to buy stamps at the post office, and so on.

A **queue** consists of a line of people or things waiting to be served and a service center with one or more servers.

For example, there would be no need of queueing in a bank if there are infinite number of people serving the customers. But that would be very expensive and impractical.

Queueing theory is applied in several disciplines such as computer systems, traffic management, operations, production, and manufacturing. It plays a significant role in modeling computer communication networks. Since the mid-1960s performance evaluation of computer communication systems are usually made using queueing models.

Reduced to its most basic form, a computer network consists of communication channels and processors (or nodes). As messages flow from node to node, queues begin to form different nodes. For high traffic intensity, the waiting or queueing time can be dominant so that the performance of the network is dictated by the behavior of the queues at the nodes. Analytical derivation of the waiting time requires a knowledge of queueing theory. Providing the basic fundamentals of queueing theory needed for the rest of the book will be our objective in this chapter.

M.N.O. Sadiku and S.M. Musa, *Performance Analysis of Computer Networks*,
DOI 10.1007/978-3-319-01646-7_4, © Springer International Publishing Switzerland 2013

4.1 Kendall's Notation

In view of the complexity of a data network, we first examine the properties of a single queue. The results from a single queue model can be extended to model a network of queues. A single queue is comprised of one or more servers and customers waiting for service. As shown in Fig. 4.1, the queue is characterized by three quantities:

- the input process,
- the service mechanism, and
- the queue discipline.

The *input process* is expressed in terms of the probability distribution of the interarrival times of arriving customers. The *service mechanism* describes the statistical properties of the service process. The *queue discipline* is the rule used to determine how the customers waiting get served. To avoid ambiguity in specifying these characteristics, a queue is usually described in terms of a well-known shorthand notation devised by D. G. Kendall [1]. In Kendall's notation, a queue is characterized by six parameters as follows:

$$\mathrm{A/B/C/K/m/z} \tag{4.1}$$

where the letters denote:

A: Arrival process, i.e. the interarrival time distribution
B: Service process, i.e. the service time distribution
C: Number of servers
K: Maximum capacity of the queue (default $= \infty$)
m: Population of customers (default $= \infty$)
z: Service discipline (default = FIFO)

The letters A and B represent the arrival and service processes and assume the following specific letters depending on which probability distribution law is adopted:

D: Constant (deterministic) law, i.e. interarrival/service times are fixed
M: Markov or exponential law, i.e. interarrival/service times are exponentially distributed
G: General law, i.e. nothing is known about the interarrival/service time distribution

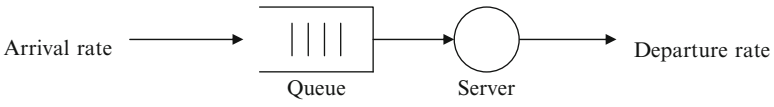

Fig. 4.1 A typical queueing system

GI: General independent law, i.e. all interarrival/service times are independent
E_k: Erlang's law of order k
H_k: Hyperexponential law of order k

The most commonly used service disciplines are:

FIFO: first-in first-out
FCFS: first-come first-serve
LIFO: last-in first-out
FIRO: first-in random-out.

It is common in practice to represent a queue by specifying only the first three symbols of Kendall's notation. In this case, it is assumed that $K = \infty$, $m = \infty$, and z = FIFO. Thus, for example, the notation M/M/1 represents a queue in which arrival times are exponentially distributed, service times are exponentially distributed, there is one server, the queue length is infinite, the customer population is infinite, and the service discipline is FIFO. In the same way, an M/G/n queue is one with Poisson arrivals, general service distribution, and n servers.

Example 4.1 A single-queue system is denoted by M/G/4/10/200/FCFS. Explain what the operation of the system is.

Solution

The system can be described as follows:

1. The interval arrival times is exponentially distributed.
2. The services times follow a general probability distribution.
3. There are four servers.
4. The buffer size of the queue is 10.
5. The population of customers to be served is 200, i.e. only 200 customers can occupy this queue.
6. The service discipline is first come, first served.

4.2 Little's Theorem

To obtain the waiting or queueing time, we apply a useful result, known as *Little's theorem* after the author of the first formal proof in 1961. The theorem relates the mean number of customers in a queue to the mean arrival rate and the mean waiting time. It states that a queueing system, with average arrival rate λ and mean waiting time per customer E(W), has a mean number of customers in the queue (or average queue length) $E(N_q)$ given by

$$E(N_q) = \lambda E(W) \tag{4.2}$$

The theorem is very general and applies to all kinds of queueing systems. It assumes that the system is in statistical equilibrium or steady state, meaning that the

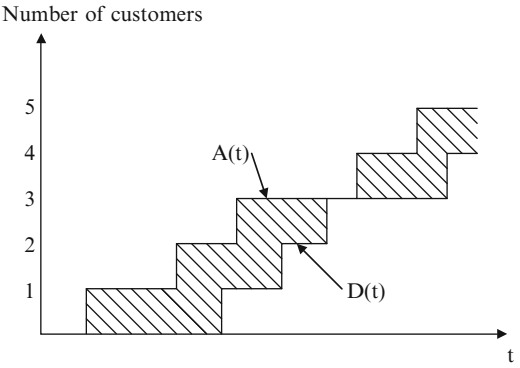

Fig. 4.2 Plot of arrival time and departure time

probabilities of the system being in a particular state have settled down and are not changing with time.

It should be noted that Eq. (4.2) is valid irrespective of the operating policies of the queueing system. For example, it holds for an arbitrary network of queues and serves. It also applies to a single queue, excluding the server.

The theorem can be proved in many ways [2–4]. Three proofs of the theorem are given by Robertazzi [2]. One of them, the graphical proof, will be given here. Suppose we keep track of arrival and departure times of individual customers for a long time t_o. If t_o is large, the number of arrivals would approximately equal to the number of departures. If this number is N_a, then

$$\text{Arrival Rate} = \lambda = \frac{N_a}{t_o} \tag{4.3}$$

Let A(t) and D(t) be respectively the number of arrivals and departures in the interval $(0,t_o)$. Figure 4.2 shows A(t) and D(t). If we subtract the departure curve from the arrival curve at each time instant, we get the number of customers in the system at that moment. The hatched area in Fig. 4.2 represents the total time spent inside the system by all customers. If this is represented by J,

$$\text{Mean time spent in system} = T = \frac{J}{N_a} \tag{4.4}$$

From Eqs. (4.3) and (4.4),

$$\text{Mean number of customers in the system} = N = \frac{J}{t_o} = \frac{N_a}{t_o} \times \frac{J}{N_a} \tag{4.5}$$

or

$$\boxed{N = \lambda T} \tag{4.6}$$

which is Little's theorem.

4.3 M/M/1 Queue

Consider the M/M/1 queue shown in Fig. 4.3.

This is a single-server system with infinite queue size, Poisson arrival process with arrival rate λ, and exponentially distributed service times with service rate μ. The queue discipline is FCFS.

The probability of k arrivals in a time interval t is given by the Poisson distribution:

$$p(k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \quad k = 0, 1, 2, \cdots \tag{4.7}$$

(Note that the Poisson arrival process has exponential arrival times.) It is readily shown that the mean or expected value and variance are given by

$$E(k) = \sum_{k=0}^{\infty} kp(k) = \lambda t \tag{4.8a}$$

$$\text{Var}(k) = E\left[(k - E(k))^2\right] = \lambda t \tag{4.8b}$$

One way of analyzing such a queue is to consider its state diagram [5–8] in Fig. 4.4.

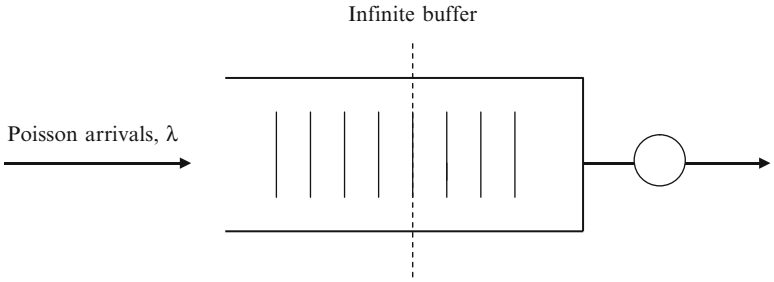

Fig. 4.3 M/M/1 queue

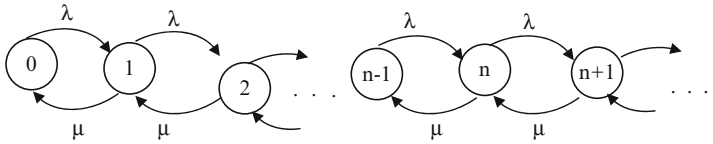

Fig. 4.4 State diagram for M/M/1 queue

We say that the system is in state n where there are n customers in the system (in the queue and the server). Notice from Fig. 4.4 that λ is the rate of moving from state n to n+1 due to an arrival in the system, whereas μ is the rate of moving from state n to n − 1 due to departure when service is completed. If N(t) is the number of customers in the system (in the queue and the server) at time t, the probability of the queue being in state n at steady state is given by

$$p_n = \lim_{t\to\infty} \text{Prob}[N(t) = n], \quad n = 0, 1, 2, \cdots \tag{4.9}$$

Our goal is to find p_n and use it to find some performance measures of interest.

Consider when the system is in state 0. Due to an arrival, the rate at which the process leaves state 0 for state 1 is λp_o. Due to a departure, the rate at which the process leaves state 1 for state 0 is μp_1. In order for stationary probability to exist, the rate of leaving state 0 must equal the rate of entering it. Thus

$$\lambda p_o = \mu p_1 \tag{4.10}$$

When the system is in state 1. Since p_1 is the proportion of time the system is in state 1, the total rate at which arrival or departure occurs is $\lambda p_1 + \mu p_1$, which is the rate at which the process leaves state 1. Similarly, the total rate at which the process enters state 1 is $\lambda p_0 + \mu p_2$. Applying the rate-equality principle gives

$$\lambda p_1 + \mu p_1 = \lambda p_0 + \mu p_2 \tag{4.11}$$

We proceed in this manner for the general case of the system being in state n and obtain

$$(\lambda + \mu)p_n = \lambda p_{n-1} + \mu p_{n+1}, \quad n \geq 1 \tag{4.12}$$

The right-hand side of this equation denotes the rate of entering state n, while the left-hand side represents the rate of leaving state n. Equations (4.10–4.12) are called *balance equations*.

We can solve Eq. (4.12) in several ways. An easy way is to write Eq. (4.12) as

$$\begin{aligned} \lambda p_n - \mu p_{n+1} &= \lambda p_{n-1} - \mu p_n \\ &= \lambda p_{n-2} - \mu p_{n-1} \\ &= \lambda p_{n-3} - \mu p_{n-2} \\ &\vdots \quad \vdots \\ &= \lambda p_0 - \mu p_1 = 0 \end{aligned} \tag{4.13}$$

Thus

$$\lambda p_n = \mu p_{n+1} \tag{4.14}$$

or

$$p_{n+1} = \rho p_n, \quad \rho = \lambda/\mu \tag{4.15}$$

If we apply this repeatedly, we get

$$p_{n+1} = \rho p_n = \rho^2 p_{n-1} = \rho^3 p_{n-2} = \cdots = \rho^{n+1} p_0, \quad n = 0, 1, 2, \cdots \tag{4.16}$$

We now apply the probability normalization condition,

$$\sum_{n=0}^{\infty} p_n = 1 \tag{4.17}$$

and obtain

$$p_0 \left[1 + \sum_{n=1}^{\infty} \rho^n \right] = 1 \tag{4.18}$$

If $\rho < 1$, we get

$$p_0 \frac{1}{1-\rho} = 1 \tag{4.19}$$

or

$$p_0 = 1 - \rho \tag{4.20}$$

From Eqs. (4.15) and (4.20),

$$\boxed{p_n = (1-\rho)\rho^n, \quad n = 1, 2, \cdots} \tag{4.21}$$

which is a geometric distribution.

Having found p_n, we are now prepared to obtain some performance measures or measures of effectiveness. These include utilization, throughput, the average queue length, and the average service time [5, 6].

The *utilization U* of the system is the fraction of time that the server is busy. In other words, U is the probability of the server being busy. Thus

$$U = \sum_{n=1}^{\infty} p_n = 1 - p_0 = \rho$$

or

$$\boxed{U = \rho} \tag{4.22}$$

The *throughput R* of the system is the rate at which customers leave the queue after service, i.e. the departure rate of the server. Thus

$$\boxed{R = \mu(1 - p_0) = \mu\rho = \lambda} \tag{4.23}$$

This should be expected because the arrival and departure rates are equal at steady state for the system to be stable.

The average number of customers in the system is

$$\begin{aligned} E(N) &= \sum_{n=0}^{\infty} np_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n \\ &= (1-\rho)\frac{\rho}{(1-\rho)^2} \end{aligned}$$

or

$$\boxed{E(N) = \frac{\rho}{1-\rho}} \tag{4.24}$$

Applying Little's formula, we obtain the *average response time* or *average delay* as

$$E(T) = \frac{E(N)}{\lambda} = \frac{1}{\lambda}\frac{\rho}{1-\rho} \tag{4.25}$$

or

$$\boxed{E(T) = \frac{1}{\mu(1-\rho)}} \tag{4.26}$$

This is the mean value of the total time spent in the system (i.e. queue and the server).

As shown in Fig. 4.5, the average delay E(T) is the sum of the average waiting time E(W) and the average service time E(S), i.e.

$$\mathrm{E(T)} = \mathrm{E(W)} + \mathrm{E(S)} \tag{4.27}$$

Equivalently, the average number of customers E(N) in the system equals the sum of the average of customers waiting $\mathrm{E(N_q)}$ in the queue and the average number of customers $\mathrm{E(N_s)}$ being served, i.e.

$$E(N) = E\left(N_q\right) + E(N_s) \tag{4.28}$$

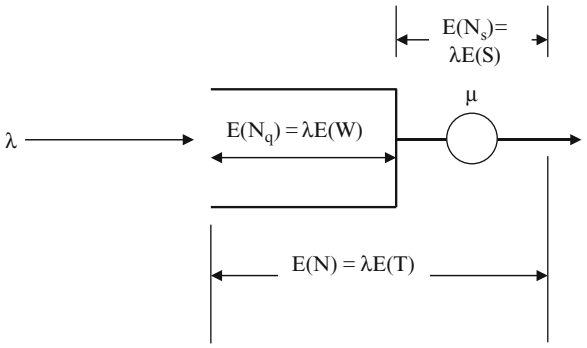

Fig. 4.5 Little's formula applied to M/M/1 queue thrice

But the mean service $E(S) = \frac{1}{\mu}$. Thus

$$E(W) = E(T) - \frac{1}{\mu} \tag{4.29}$$

or

$$\boxed{E(W) = \frac{\rho}{\mu(1-\rho)}} \tag{4.30}$$

We now apply Little's theorem to find the *average queue length* or the average number of customers waiting in the queue, i.e.

$$\boxed{E(N_q) = \lambda E(W) = \frac{\rho^2}{1-\rho}} \tag{4.31}$$

Finally, since $E(N) = \lambda E(T)$, it is evident from Eqs. (4.27) and (4.28) that

$$E(N_s) = \lambda E(S) = \lambda \frac{1}{\mu} = \rho \tag{4.32}$$

Notice from Eqs. (4.25), (4.31), (4.32) that the Little's theorem is applied three times. This is also shown in Fig. 4.5.

Example 4.2 Service at a bank may be modeled as an M/M/1 queue at which customers arrive according to Poisson process. Assume that the mean arrival rate is 1 customer/min and that the service times are exponentially distributed with mean 40 s/customer. (a) Find the average queue length. (b) How long does a customer have to wait in line? (c) Determine the average queue size and the waiting time in the queue if the service time is increased to 50 s/customer.

Solution

As an M/M/1 queue, we obtain mean arrival rate as
$\lambda = 1$ customer/min
and the mean service rate as

$$E(S) = \frac{1}{\mu} = 40\,\text{s/customer} = \frac{40}{60}\,\text{min/customer}$$

Hence, the traffic intensity is

$$\rho = \frac{\lambda}{\mu} = (1)(40/60) = \frac{2}{3}$$

(a) The mean queue size is

$$E[N_q] = \frac{\rho^2}{1-\rho} = \frac{(2/3)^2}{1-2/3} = 1.333 \text{ customers}$$

(b) The mean waiting time is

$$E[W] = \frac{\rho}{\mu(1-\rho)} = \frac{2/3(4/6)}{(1-2/3)} = 1.333\,\text{min}$$

(c) If the mean service time E(S) = 50 s/customer = 50/60 min/customer, then

$$\rho = \frac{\lambda}{\mu} = (1)(50/60) = \frac{5}{6}$$

$$E[N_q] = \frac{\rho^2}{1-\rho} = \frac{(5/6)^2}{1-5/6} = 4.1667 \text{ customers}$$

$$E[W] = \frac{\rho}{\mu(1-\rho)} = \frac{5/6(5/6)}{(1-5/6)} = 4.1667\,\text{min}$$

We expect the queue size and waiting time to increase if it takes longer time for customers to be served.

4.4 M/M/1 Queue with Bulk Arrivals/Service

In the previous section, it was assumed that customers arrive individually (or one at a time) and are provided service individually. In this section, we consider the possibility of customers arriving in bulk (or in groups or batch) or being served in bulk. Bulk arrivals/service occur in practice because it is often more economical to collect a number of items (jobs, orders, etc.) before servicing them.

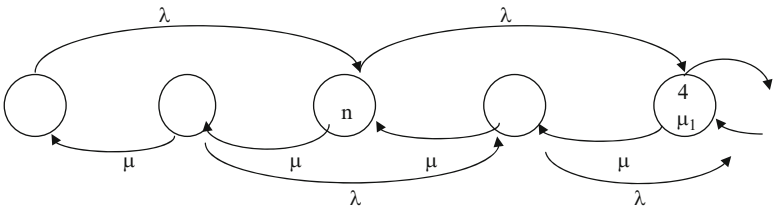

Fig. 4.6 Transition diagram of $M^X/M/1$ queue with m = 2

4.4.1 $M^x/M/1$ (Bulk Arrivals) System

Here we consider the situation where arrivals occur in batches of more than one customer, i.e. in bulk. Although the process is not birth-and-death process, the arrival instants still occur as a Poisson process with constant rate λ. Each of the arriving customers is served in standard fashion (first-come, first served, one at a time) by a server with exponentially distributed service times with parameter μ. Suppose the size of the batch is fixed at $m \geq 1$ customers. Then only two transitions can occur as

$$n \rightarrow n + m \quad \text{(arrival)}$$

or

$$n + 1 \rightarrow n \quad \text{(departure)}$$

The state transition diagram is shown in Fig. 4.6 for m = 2.

The balance equation for n = 0 is

$$\lambda p_0 = m\mu p_1 \tag{4.33}$$

and for $n \geq 1$ is

$$(\lambda + \mu m)p_n = \mu m p_{n+1} + \lambda p_{n-m} \tag{4.34}$$

We now apply the method of z-transforms to solve for p_n. We define the generating function

$$G(z) = \sum_{i=0}^{\infty} p_n z^n \tag{4.35}$$

Multiplying the balance equation for state n by z^n and summing, we obtain

$$\sum_{n=1}^{\infty} (\lambda + \mu m)p_n z^n = \sum_{n=1}^{\infty} \mu m p_{n+1} z^n + \sum_{n=1}^{\infty} \lambda p_{n-m} z^n \tag{4.36}$$

Simplifying yields

$$G(z) = \frac{\mu m(1-z)p_0}{\mu m + \lambda z^{m+1} - z(\lambda + \mu m)} \tag{4.37}$$

The value of p_0 is obtained using the condition $G(1) = 1$.

$$p_0 = 1 - \frac{\lambda m}{\mu} = 1 - \rho, \quad \rho = \frac{\lambda m}{\mu} \tag{4.38}$$

4.4.2 $M/M^Y/1$ (Bulk Service) System

This kind of model is used to analyze systems that wait until a certain message size is reached before releasing the data for transmission. We will assume that customers are served in bulk of size m, i.e. customers are served m at a time. At equilibrium, the balance equations are [8, 9]:

$$(\lambda + \mu)p_n = \lambda p_{n-1} + \mu p_{n+m}, \quad n \geq 1 \tag{4.39a}$$

$$\lambda p_0 = \mu p_m + \mu p_{m-1} + \cdots + \mu p_1 \tag{4.39b}$$

Equation (4.39a) can be written in terms of an operator D so

$$\left[\mu D^{m+1} - (\lambda + \mu)D + \lambda\right]p_n = 0, \quad n \geq 0 \tag{4.40}$$

If the roots of the characteristic equation are $r_1, r_2, \cdots, r_{m+1}$, then

$$p_n = \sum_{I=1}^{m+1} C_i r_i^n, \quad n \geq 0 \tag{4.41}$$

Using the fact that $\sum_{n=0}^{\infty} p_n = 1$, we obtain

$$p_n = (1 - r_0)r_0^n, \quad n \geq 0, 0 < r_0 < 1 \tag{4.42}$$

where r_o is the one and only one root of Eq. (4.40) that is less than one. Comparing this with Eq. (4.21) shows the similarity between this solution and that of M/M/1. Hence,

$$E[N] = \frac{r_0}{1 - r_0} \tag{4.43}$$

$$E[T] = \frac{r_0}{\lambda(1 - r_0)} \tag{4.44}$$

4.5 M/M/1/k Queueing System

In this case, we have situations similar to M/M/1 but the number of customers that can be queued is limited to k. In other words, this is a system with limited waiting space. If an arriving customer finds the queue full, it is lost or blocked, as shown in Fig. 4.7.

Hence,

$$\lambda_n = \begin{cases} \lambda, & \text{if } 0 \le n < k \\ 0, & n \ge k \end{cases} \tag{4.45}$$

$$\mu_n = \mu, \quad 0 \le n \le k \tag{4.46}$$

The state transition diagram is given in Fig. 4.8.

The balance equations are

$$\begin{gathered} \lambda p_0 = \mu p_1 \\ \lambda p_n + \mu p_n = \lambda p_{n-1} + \mu p_{n+1}, \quad 1 \le n \le k-1 \\ \lambda p_{k-1} = \mu p_k \end{gathered} \tag{4.47}$$

We solve these equations recursively and apply the normalization condition. If we define $\rho = \lambda/\mu$, the state probabilities at steady state are given by

$$p_n = \begin{cases} \dfrac{(1-\rho)\rho^n}{1-\rho^{k+1}}, & 0 \le n \le k \\ 0, & n > k \end{cases} \tag{4.48}$$

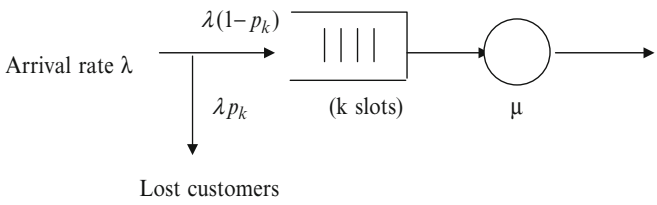

Fig. 4.7 M/M/1/k queueing system

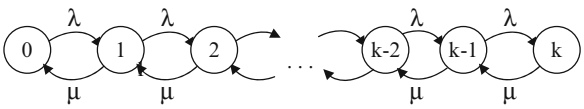

Fig. 4.8 State transition diagram for the M/M/1/k queue

The utilization of the server is given by

$$U = 1 - p_0 = \frac{\rho\left(1-\rho^k\right)}{1-\rho^{k+1}} \tag{4.49}$$

The average queue length is

$$E\left(N_q\right) = \sum_{n=0}^{k} np_n = \frac{\rho}{1-\rho^{k+1}}\left[\frac{1-\rho^k}{1-\rho} - k\rho^k\right] \tag{4.50}$$

Since there can be blocking in this system, the blocking probability is

$$\boxed{P_B = p_k = \frac{(1-\rho)\rho^k}{1-\rho^{k+1}}} \tag{4.51}$$

This is the probability that arriving customer is blocked, i.e. it is lost because it finds the queue full.

Example 4.3 A system consists of a packet buffer and a communication server and can hold not more than three packets. Arrivals are Poisson with rate 15 packets/ms and the server follows exponential distribution with mean 30 packets/ms. Determine the blocking probability of the system.

Solution

This is an M/M/1/k system with k = 3.

$$\rho = \lambda\frac{1}{\mu} = \frac{15}{30} = 0.5$$

The probability is

$$P_B = \frac{(1-\rho)\rho^k}{1-\rho^{k+1}} = \frac{(1-0.5)0.5^3}{1-0.5^4} = 0.0667$$

which is about 7 %.

4.6 M/M/k Queueing System

This is the case where we have k servers, as shown in Fig. 4.9.

Upon arrival, a customer is served by any available server. The arriving customer is queued when all servers are found busy, i.e. no customer is queued until the number of arrivals exceeds k. The state transition diagram is shown in Fig. 4.10.

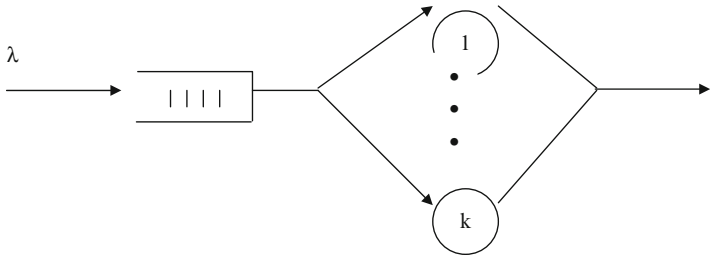

Fig. 4.9 The M/M/k queue

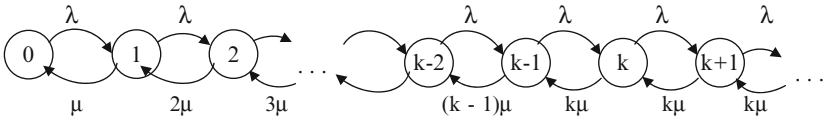

Fig. 4.10 State transition diagram for M/M/k system

The system can be modeled as a birth-and-death process with

$$\lambda_n = \lambda \tag{4.52}$$

$$\mu_m = \begin{cases} n\mu, & 0 \le n \le k \\ k\mu, & n \ge k \end{cases}$$

At steady state,

$$\lambda p_{n-1} = n\mu p_n, \quad n \le k \tag{4.53a}$$

$$\lambda p_{n-1} = k\mu p_n, \quad n > k \tag{4.53b}$$

From these, we obtain the state probabilities as

$$p_n = \begin{cases} p_0 \dfrac{(k\rho)^n}{n!}, & n \le k \\ p_0 \dfrac{\rho^n k^k}{k!}, & n \ge k \end{cases} \tag{4.54}$$

where $\rho = \frac{\lambda}{k\mu} < 1$. Solving for p_0, we get

$$p_0 = \left[\sum_{n=0}^{k-1} \frac{(k\rho)^n}{n!} + \left(\frac{k^k \rho^k}{k!} \right) \frac{1}{1-\rho} \right]^{-1} \tag{4.55}$$

Measures of effectiveness for this model can be obtained in the usual manner. The probability that an arriving customer joins the queue is

$$\text{Prob[queueing]} = P_Q = \sum_{n=k}^{\infty} p_n = \sum_{n=k}^{\infty} \frac{p_0 k^k \rho^n}{k!} = \frac{p_0 (k\rho)^k}{k!} \sum_{n=k} \rho^{n-k} = \frac{k^k \rho^k}{k!} \left(\frac{p_0}{1-\rho} \right)$$

or

$$\boxed{P_Q = \frac{k^k \rho^k}{k!} \left(\frac{p_0}{1-\rho} \right)} \tag{4.56}$$

This formula is known as Erlang's C formula. It is widely used in telephony; it gives the probability that no trunk (or server) is available for an arriving call.

The average queue length is

$$E[N] = \sum_{n=0}^{\infty} n p_n = k\rho + \frac{\rho}{(1-\rho)} P_Q \tag{4.57}$$

Using Little's theorem, the average time spent E[T] in the system can be obtained as

$$E[T] = \frac{E[N]}{\lambda} = \frac{1}{\mu} + \frac{1}{\mu k} \frac{P_Q}{(1-\rho)} \tag{4.58}$$

4.7 M/M/∞ Queueing System

This is the case in which we have infinite number of servers so that an arriving customer can always find a server and need not queue This model can be used to study the effect of delay in large systems. The state transition diagram for the M/M/∞ system is shown in Fig. 4.11.

Like we did before, we assume a Poisson arrivals at rate λ and exponentially distributed service times with mean 1/μ. We adopt a birth-and-death process with parameters

$$\lambda_n = \lambda, \quad n = 0, 1, 2, \cdots \tag{4.59}$$

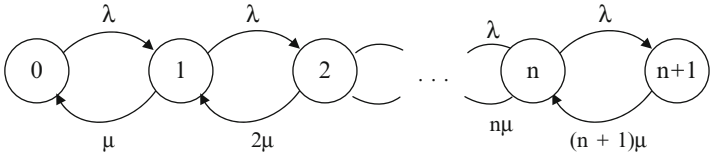

Fig. 4.11 State transition diagram for M/M/∞ queueing system

$$\mu_n = n\mu, \quad n = 1, 2, \cdots \tag{4.60}$$

The balance equation is

$$\lambda p_n = (n+1)\mu p_{n+1} \tag{4.61}$$

which can be solved to give

$$p_n = \frac{\rho^n}{n!} p_0 \tag{4.62}$$

where $\rho = \lambda/\mu$. Applying the normalization condition $\sum_{n=0}^{\infty} p_n = 1$ gives

$$p_0 = e^{-\rho} \tag{4.63}$$

The utilization of the server is

$$U = 1 - p_0 = 1 - e^{-\rho} \tag{4.64}$$

The average number of customers in the system is

$$E[N] = \sum_{n=0}^{\infty} n p_n = \rho \tag{4.65}$$

We apply Little's theorem in finding the average time spent in the system.

$$E[T] = \frac{E[N]}{\lambda} = \frac{1}{\mu} \tag{4.66}$$

Also,

$$E\left[N_q\right] = 0 = E\left[W_q\right] \tag{4.67}$$

i.e. the average waiting time and the average number of customers waiting in the queue are both zero.

4.8 M/G/1 Queueing System

The M/G/1 queueing system is the simplest non-Markovian system. We analyze it assuming that it is in the steady state. An M/G/1 system assumes a FIFO service discipline, an infinite queue size, a Poisson input process (with arrival rate λ), a general service times (with arbitrary but known distribution function H, mean

$\tau = 1/\mu$, and variance σ^2), and one server. To derive the average waiting time of the M/G/1 model requires some effort beyond the scope of this book. The derivation involves applying the *method of z-transform* or generating functions and is provided in the Appendix A for the curious student. The result is [10–12]:

$$\boxed{E(W) = \frac{\rho\tau}{2(1-\rho)}\left(1 + \frac{\sigma^2}{\tau^2}\right)} \tag{4.68}$$

where $\rho = \lambda/\mu = \lambda\tau$. This is known as *Pollaczek-Khintchine formula* after two Russian mathematicians Pollaczek and Khintchine who derived the formula independently in 1930 and 1932 respectively. The average number of customers $E(N_q)$ in the queue is

$$E(N_q) = \lambda E(W) = \frac{\rho^2}{2(1-\rho)}\left(1 + \frac{\sigma^2}{\tau^2}\right) \tag{4.69}$$

The average response time is

$$E(T) = E(W) + \tau = \frac{\rho\tau}{2(1-\rho)}\left(1 + \frac{\sigma^2}{\tau^2}\right) + \tau \tag{4.70}$$

and the mean number of customers in the system is

$$E(N) = \lambda E(T) = E(N_q) + \rho \tag{4.71}$$

or

$$E(N) = \frac{\rho^2}{2(1-\rho)}\left(1 + \frac{\sigma^2}{\tau^2}\right) + \rho \tag{4.72}$$

We may now obtain the mean waiting time for the M/M/1 and M/D/1 queue models as special cases of the M/G/1 model.

For the M/M/1 queue model, a special case of the M/G/1 model, the service times follow an exponential distribution with mean $\tau = 1/\mu$ and variance σ^2. That means,

$$H(t) = \text{Prob}[X \leq t] = 1 - e^{-\mu t} \tag{4.73}$$

Hence,

$$\sigma^2 = \tau^2 \tag{4.74}$$

Substituting this in Pollaczek-Khintchine formula in Eq. (4.68) gives the mean waiting time as

$$E(W) = \frac{\rho\tau}{(1-\rho)} \tag{4.75}$$

The M/D/1 queue is another special case of the M/G/1 model. For this model, the service times are constant with the mean value $\tau = 1/\mu$ and variance $\sigma = 0$. Thus Pollaczek-Khintchine formula in Eq. (4.68) gives the mean waiting time as

$$E(W) = \frac{\rho\tau}{2(1-\rho)} \tag{4.76}$$

It should be noted from Eqs. (4.75) and (4.76) that the waiting time for the M/D/1 model is one-half that for the M/M/1 model, i.e.

$$E(W)_{M/D/1} = \frac{\rho\tau}{2(1-\rho)} = \frac{1}{2}E(W)_{M/M/1} \tag{4.77}$$

Example 4.4 In the M/G/1 system, prove that:

(a) Prob (the system is empty) = 1 − ρ
(b) Average length of time between busy periods = 1/λ
(c) Average no. of customers served in a busy period = $\frac{1}{1-\rho}$

where $\rho = \lambda\overline{X}$ and $\overline{X}$ is the mean service time.

Solution

(a) Let p_b = Prob. that the system is busy. Then p_b is the fraction of time that the server is busy. At steady state, arrival rate = departure rate

$$\lambda = p_b\mu$$

or

$$p_b = \frac{\lambda}{\mu} = \rho$$

The Prob. that the system is empty is

$$p_e = 1 - p_b = 1 - \rho$$

(b) The server is busy only when there are arrivals. Hence the average length of time between busy periods = average interarrival rate = 1/λ.

Alternatively, we recall that if t is the interarrival time,

$$f(t) = \lambda e^{-\lambda t}$$

Hence E(t) = 1/λ.

(c) Let E(B) = average busy period, E(I) = average idle period. From part (a),

$$p_b = \rho = \frac{E(B)}{E(B) + E(I)}$$

From part (b),
E(I) = average length of time between busy periods = 1/λ
Hence

$$\rho = \frac{E(B)}{E(B) + \frac{1}{\lambda}}$$

Solving for E(B) yields

$$E(B) = \frac{\rho}{\lambda(1-\rho)} = \frac{\overline{X}}{1-\rho}$$

as required.
The average no. of customers served in a busy period is

$$N_b = \frac{\text{Average length of busy period}}{\text{Average service time}}$$

Hence

$$N_b = E(B)/\overline{X} = \frac{1}{1-\rho}$$

4.9 M/E_k/1 Queueing System

In this case, the service time distribution is Erlang distribution with parameters μ and k, i.e.

$$f_X(x) = \frac{\mu(\mu x)^{k-1}}{(k-1)!} e^{-\mu x}, \quad x \geq 0 \tag{4.78}$$

with mean and variance

$$E[X] = \frac{k}{\mu}, \quad \text{Var}[X] = \frac{k}{\mu^2} \tag{4.79}$$

This should be regarded as another special case of M/G/1 system so that Pollaczek-Khintchine formula in Eq. (4.68) applies. Thus,

$$E\left[W_q\right] = \frac{1+k}{2k} \frac{\lambda}{\mu(\mu-\lambda)} = \frac{1+k}{2k} \frac{\rho}{\mu(1-\rho)} \tag{4.80}$$

$$E\left[N_q\right] = \lambda E\left(W_q\right) = \frac{1+k}{2k} \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{1+k}{2k} \frac{\rho^2}{1-\rho} \tag{4.81}$$

$$E[T] = E[W_q] + \frac{1}{\mu} \tag{4.82}$$

$$E[N] = \lambda E[T] \tag{4.83}$$

where $\rho = \lambda/\mu$.

4.10 Networks of Queues

The queues we have considered so far are isolated. In real life, we have a network of queues interconnected such as shown in Fig. 4.12. Such networks of queues are usually complicated and are best analyzed using simulation. However, we consider two simple ones here [13–15].

4.10.1 Tandem Queues

Consider two M/M/1 queues in tandem, as shown in Fig. 4.13. This is an example of open queueing network.

The state diagram is shown in Fig. 4.14. From the sate diagram, we can obtain the balance equations.

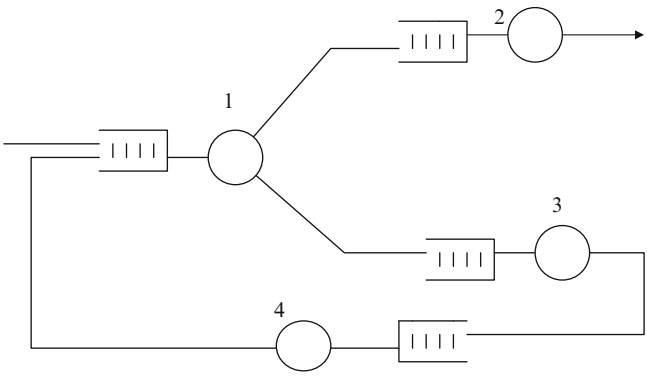

Fig. 4.12 A typical network of queues

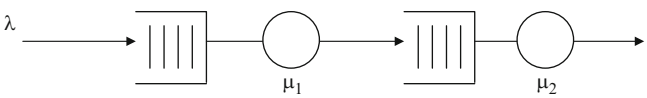

Fig. 4.13 Two M/M/1 queues in tandem

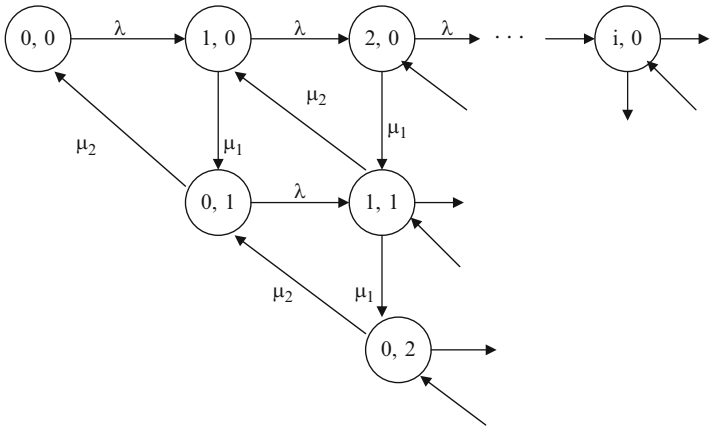

Fig. 4.14 The state diagram for two M/M/1 queues in tandem

Let

$$p_{i,j} = \text{Prob}[\text{ i jobs at server 1 and j jobs at server 2}]$$

For state (0,0),

$$\lambda p_{0,0} = \mu_2 p_{0,1} \tag{4.84}$$

For state (i,0), $i > 0$,

$$\lambda p_{i-1,0} + \mu_2 p_{i,1} - (\lambda + \mu_1) p_{i,0} = 0 \tag{4.85}$$

For state (0,j), $j > 0$,

$$\mu_1 p_{1,j-1} + \mu_2 p_{0,j+1} - (\lambda + \mu_2) p_{0,j} = 0 \tag{4.86}$$

For state (i,j),

$$\lambda p_{i-1,j} + \mu_1 p_{i+1,j-1} + \mu_2 p_{i,j+1} - (\lambda + \mu_1 + \mu_2) p_{i,j} = 0 \tag{4.87}$$

Since queue 1 is unaffected by what happens at queue 2, the marginal probability of i jobs at queue 1 is

$$p_i = (1 - \rho_1)\rho_1^i, \quad \rho_1 = \frac{\lambda}{\mu_1} \tag{4.88}$$

Similarly, for queue 2

$$p_j = (1 - \rho_2)\rho_2^j, \quad \rho_2 = \frac{\lambda}{\mu_2} \tag{4.89}$$

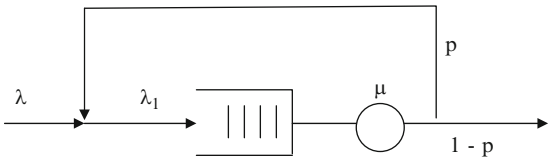

Fig. 4.15 A queueing system with a (Bernoulli) feedback

A simple product form solution for this two-node network is

$$p_{i,j} = (1-\rho_1)(1-\rho_2)\rho_1^i \rho_2^j, \quad \rho_1 = \frac{\lambda}{\mu_1} \tag{4.90}$$

The analysis of even this simplest case is extremely complicated.

4.10.2 Queueing System with Feedback

Queuing systems with feedback are applicable to a fairly limited set of circumstances. A typical example is shown in Fig. 4.15. The problem here is that the combination of the external Poisson process and the feedback process is not Poisson because the processes being superposed are not independent due to the feedback. However, consideration of the steady state diagram shows us that, as far as queue length is concerned, the system behaves like an M/M/1 queue with arrival rate λ and service rate pμ. Also, the traffic equation for this network is

$$\lambda_1 = \lambda + \lambda_1 p \rightarrow \lambda_1 = \frac{\lambda}{1-p} \tag{4.91}$$

4.11 Jackson Networks

A Jackson network has a steady state solution in product form. Such product-form queueing networks can be open or closed. The nature of such networks allows us to decouple the queues, analyze them separately as individual systems, and then combine the results. For example, consider a series of k single-server queues with exponential service time and Poisson arrivals, as shown in Fig. 4.16.

Customers entering the system join queue at each stage. It can be shown that each queue can be analyzed independently of other queues. Each queue has an arrival and a departure rate of λ. If the ith server has a service rate of μ_i, the utilization of the ith server is

$$\rho_i = \frac{\lambda}{\mu_i} \tag{4.92}$$

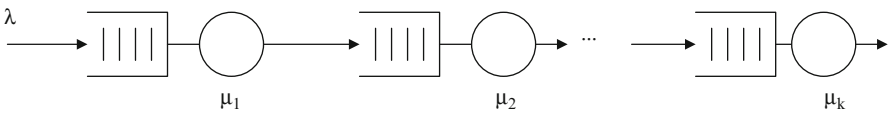

Fig. 4.16 k M/M/1 queues in series

and

$$\text{Prob}[n_i \text{ customers in the ith queue}] = P(n_i) = (1-\rho_i)\rho_i^{n_i} \tag{4.93}$$

The joint probability of queue lengths of k queues is the product of individual probabilities.

$$\begin{aligned} P(n_1, n_2, \cdots, n_k) &= (1-\rho_1)\rho_1^{n_1}(1-\rho_2)\rho_2^{n_2}\cdots(1-\rho_k)\rho_k^{n_k} \\ &= P_1(n_1)P_2(n_2)\cdots P_k(n_k) \end{aligned} \tag{4.94}$$

This is known as *Jackson theorem*, after J.R. Jackson who first proved the property. The queueing network is therefore a product-form network. A network to which Jackson's theorem is applicable is known as *Jackson network*. In general, for a product-form network

$$P(n_1, n_2, \cdots, n_k) = \frac{1}{G}\prod_{i=1}^{k}\rho_i^{n_i} \tag{4.95}$$

where G is a normalization constant and is a function of the total number of jobs in the system. The product-form networks are easier to solve than nonproduct-form networks.

4.12 Summary

1. A simple introduction to queueing theory was presented.
2. Beginning with the M/M/1 queue, we derived the closed form expressions for some performance measures.
3. We also considered the case of an M/M/1 queue with bulk arrivals or service. We considered M/M/1/k, M/M/k, and M/M/∞ queueing systems.
4. Using the more general queueing model M/G/1, we derived the Pollaczek-Khintchine formula for the mean waiting time. The corresponding mean waiting times for the M/M/1, M/D/1, M/E_k/1 queue models were derived as special cases of the M/G/1 model.

A more in depth introduction to queueing theory can be found in [11, 12, 16–22]. We will apply the ideas in this chapter to model computer networks in the following chapters.

Problems

4.1 For the M/M/1 system, find: (a) $E(N^2)$, (b) $E(N(N-1))$, (c) Var(N).

4.2 In an M/M/1 queue, show that the probability that the number of messages waiting in the queue is greater than a certain number m is

$$P(n > m) = \rho^{m+1}$$

4.3 For an M/M/1 model, what effect will doubling λ and μ have on E[N], $E[N_q]$, and E[W]?

4.4 Customers arrive at a post office according to a Poisson process with 20 customers/h. There is only one clerk on duty. Customers have exponential distribution of service times with mean of 2 min. (a) What is the average number of customers in the post office? (b) What is the probability that an arriving customer finds the clerk idle?

4.5 From the balance equation for the M/M/1 queue, obtain the probability generating function.

4.6 An air-line check-in counter at Philadelphia airport can be modeled as an M/M/1 queue. Passengers arrive at the rate of 7.5 customers per hour and the service takes 6 min on the average. (a) Find the probability that there are fewer than four passengers in the system. (b) On the average, how long does each passenger stay in the system? (c) On the average, how many passengers need to wait?

4.7 An observation is made of a group of telephone subscribers. During the 2-h observation, 40 calls are made with a total conversation time of 90 min. Calculate the traffic intensity and call arrival rate assuming M/M/1 system.

4.8 Customers arrive at a bank at the rate of 1/3 customer per minute. If X denotes the number of customers to arrive in the next 9 min, calculate the probability that: (a) there will be no customers within that period, (b) exactly three customers will arrive in this period, and (c) at least four customers will arrive. Assume this is a Poisson process.

4.9 At a telephone booth, the mean duration of phone conversation is 4 min. If no more than 2-min mean waiting time for the phone can be tolerated, what is the mean rate of the incoming traffic that the phone can support?

4.10 For an M/M/1 queue operating at fixed $\rho = 0.75$, answer the following questions: (a) Calculate the probability that an arriving customer finds the queue empty. (b) What is the average number of messages stored? (c) What is the average number of messages in service? (d) Is there a single time at which this average number is in service?

4.11 At a certain hotel, a lady serves at a counter and she is the only one on duty. Arrivals to the counter seem to follow the Poisson distribution with mean of 10 customers/h. Each customer is served one at a time and the service time follows an exponential distribution with a mean of 4 min.

(a) What is the probability of having a queue?
(b) What is the average queue length?
(c) What is the average time a customer spends in the system?
(d) What is the probability of a customer spending more than 5 min in the queue before being attended to?

Note that the waiting time distribution for an M/M/1 queue is

$$\text{Prob}(W > t) = W(t) = 1 - \rho e^{-\mu(1-\rho)t}, \quad t \geq 0$$

4.12 (a) The probability p_n that an infinite M/M/2 queue is in state n is given by

$$p_n = \begin{cases} \dfrac{(1-\rho)}{(1+\rho)}, & n = 0 \\ \dfrac{2(1-\rho)}{(1+\rho)}\rho^2, & n \geq 0 \end{cases}$$

where $\rho = \frac{\lambda}{2\mu}$. Find the average occupancy E(N) and the average time delay in the queue E(T).

4.13 Consider M/M/k model. Show that the probability of any server is busy is $\lambda / k\mu$.

4.14 For the M/M/1/k system, let q_n be the probability that an arriving customer finds n customers in the system. Prove that

$$q_n = \frac{p_n}{1 - p_k}$$

4.15 Derive Eq.(4.62) from Eq. (4.61).

4.16 Find the mean and variance of the number of customers in the system for the M/M/∞ queue.

4.17 At a toll booth, there is only one "bucket" where each driver drops 25 cents. Assuming that cars arrive according to a Poisson probability distribution at rate 2 cars per minute and that each car takes a *fixed* time 15 s to service, find: (a) the long-run fraction of time that the system is busy, (b) the average waiting time for each car, (c) the average number of waiting cars, (d) how much money is collected in 2 h.

4.18 An $M/E_k/1$ queue has an arrival rate of 8 customers/s and a service rate of 12 customers/s. Assuming that k = 2, find the mean waiting time.

4.19 Consider two identical M/M/1 queueing systems in operation side by side in a facility with the same rates λ and μ ($\rho = \lambda/\mu$). Show that the distribution of the total number N of customers in the two systems combined is

$$\text{Prob}(N = n) = (n+1)(1-\rho)^2\rho^n, \quad n > 0$$

References

1. D. G. Kendall, "Some problems in the theory of queues," *J. Roy. Statis. Soc*. Series B, vol. 13, 1951, pp. 151–185.
2. T. G. Robertazzi, *Computer Networks and Systems: Queueing Theory and Performance Evaluation*. New York: Springer-Verlag, 1990, pp. 43-47.
3. S. Eilon, "A Simpler Proof of L = λW," *Operation Research*, vol. 17, 1969, pp. 915–916.
4. R. Jain, *The Art of Computer Systems Performance Analysis*. New York: John Wiley, 1991, pp. 513-514.
5. J. Medhi, *Stochastic Models in Queueing Theory*. San Diego, CA: Academic Press, 1991, pp. 71–75.
6. G. C. Cassandras, *Discrete Event Systems*. Boston, MA: Irwin, 1993, pp.349-354, 404-413.
7. M. Schartz, *Telecommunication Networks*. Reading, MA: Addison-Wesley, 1987, pp. 21-69.
8. D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*. New York: John Wiley, 1998, 3rd ed., pp. 116-164.
9. E. Gelenbe and G. Pujolle, *Introduction to Queueing Networks*. Chichester, UK: John Wiley & Sons, 1987, pp. 94-95.
10. R. Nelson, *Probability, Stochastic Processes, and Queueing Theory*. New York: Springer-Verlag, 1995, pp. 295–309.
11. R. B. Cooper, *Introduction to Queueing Theory*. New York: North-Holland, 2nd ed., 1981, pp. 208-222.
12. R. B. Cooper, "Queueing Theory," in D. P. Heyman (ed.), *Handbooks in Operations Research and Management Science*. New York: North-Holland, 1990, chap. 10, pp. 469-518.
13. P. J.B. King, *Computer and Communication System Performancd Modelling*. New York: Prentice Hall,1989.pp.124-130
14. P. G. Harrison and N. M. Patel, *Performance Modelling of Communication Networks and Computer Architecture*. Wokingham, UK: Addison-Wesley, 1993, pp. 258-297.
15. M. K. Molloy, *Fundamentals of Performance Modeling*. New York: MacMillan, 1989, pp. 193-248.
16. L. Kleinrock, *Queueing Systems*. New York: John Wiley, 1975, vol. I.
17. J. D. Claiborne, *Mathematical Preliminaries for Computer Networking*. New York: John Wiley, 1990.
18. O. C. Ibe, *Markov Processes for Stochastic Modeling*. Burlington, MA: Elsevier Academic Press, 2009, pp. 105-152.
19. —, *Fundamentals of Stochastic Networks*. New York: John Wiley & Sons, 2011.
20. J. F. Hayes and T. V. J. G. Babu, *Modeling and Analysis of Telecommunications Networks*. New York: Wiley-Interscience, 2004, pp. 67-112.
21. A. M. Haghighi and D. P. Mishev, *Queueing Models in Industry and Business*. New York: Nova Science Publishers, 2008.
22. G. R. Dattatreya, *Performance Analysis of Queuing and Computer Networks*. Boca Raton, FL: CRC Press, 2008.