# Queueing Theory

## INTRODUCTION

Queueing theory deals with the study of queues (waiting lines). Queues abound in practical situations. The earliest use of queueing theory was in the design of a telephone system. Applications of queueing theory are found in fields as seemingly diverse as traffic control, hospital management, and time-shared computer system design. In this chapter, we present an elementary queueing theory.

## QUEUEING SYSTEMS

### A. Description:

A simple queueing system is shown in Fig. 16-1. Customers arrive randomly at an average rate of $\lambda_a$ **.** Upon arrival, they are served without delay if there are available servers; otherwise, they are made to wait in the queue until it is their turn to be served. Once served, they are assumed to leave the system. We will be interested in determining such quantities as the average number of customers in the system, the average time a customer spends in the system, the average time spent waiting in the queue, etc.

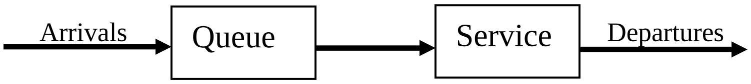


Fig.16-1 A simple queuing system

The description of any queueing system requires the specification of three parts:

**1. The arrival process**
**2. The service mechanism, such as the number of servers and service-time distribution**
**3. The queue discipline (for example, first-come, first-served)**

### B. Classification :

The notation ***A/B/s/K*** is used to classify a queueing system, where
***A*** specifies the type of arrival process,
***B*** denotes the service-time distribution,
***s*** specifies the number of servers,
and ***K*** denotes the capacity of the system, that is, the maximum number of customers that can be accommodated.
If K is not specified, it is assumed that the capacity of the system is unlimited.

**Examples:**

M/M/2 queueing system (M stands for Markov) is one with Poisson arrivals, exponential service-time distribution, and 2 servers.

An **M/G/l** queueing system has Poisson arrivals, general service-time distribution, and a single server.

A special case is the M/D/1 queueing system, where D stands for constant (deterministic:) service time.

**Examples of queueing**

systems with limited capacity are telephone systems with limited trunks, hospital emergency rooms with limited beds, and airline terminals with limited space in which to park aircraft for loading and unloading. In each case, customers who arrive when the system is saturated are denied entrance and are lost.

## C. Useful Formulas

Some basic quantities of queueing systems are
L: the average number of customers in the system
$L_q$: the average number of customers waiting in the queue
$L_s$: the average number of customers in service
*W:* the average amount of time that a customer spends in the system
$W_q$: the average amount of time that a customer spends waiting in the queue
$W_s$: the average amount of time that a customer spends in service

Many useful relationships between the above and other quantities of interest can be obtained by using the following basic cost identity:

Assume that entering customers are required to pay an entrance fee (according to some rule) to the system. Then we have:
Average rate at which the system earns = $\lambda_a$ x average amount an entering customer pays (16.1)
where $\lambda_a$ **,** is the average arrival rate of entering customers

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

and X(t) denotes the number of customer arrivals by time t.
If we assume that each customer pays $1 per unit time while in the system , then Eq. 16.1 yeilds:

$$\mathbf{L} = \lambda_a \ \mathbf{x} \ \mathbf{W} \qquad (16.2)$$

Equation (16.2) is sometimes known as Little's formula.
Similarly, if we assume that each customer pays $1 per unit time while in the queue,then Eq. 16.1 yields

$$\text{Lq} = \lambda_a \ \text{x}\, W_q \qquad (16.3)$$

If we assume that each customer pays $1 per unit time while in service, Eq. (16.1) yields

$$\text{Ls} = \lambda_a \ \text{x}\, W_S \qquad (16.4)$$

**Note that Eqs. (16.2) to (16.4) are valid for almost all queueing systems, regardless of the arrival process, the number of servers, or queueing discipline.**

**BIRTH-DEATH PROCESS**
We say that the queueing system is in state $S_n$**,** if there are ***n*** customers in the system, including those being served. Let N(t) be the Markov process that takes on the value n when the queueing system is in state $S_n$, with the following assumptions:

1. If the system is in state $S_n$, it can make transitions only to $S_{n-1}$, or $S_{n+1}$ **,** , $n \geq 1$ ; that is, either a customer completes service and leaves the system or, while the present customer is still being serviced, another customer arrives at the system ; from $S_o$ , the next state can only be $S_1$ **.**

2. If the system is in state $S_n$, at time t, the probability of a transition to $S_{n+1}$, in the time interval (t, t $+ \ \Delta$ t) is $\mathbf{a_n} \ \Delta$ t. We refer to $a_n$ as the arrival parameter (or the birth parameter).

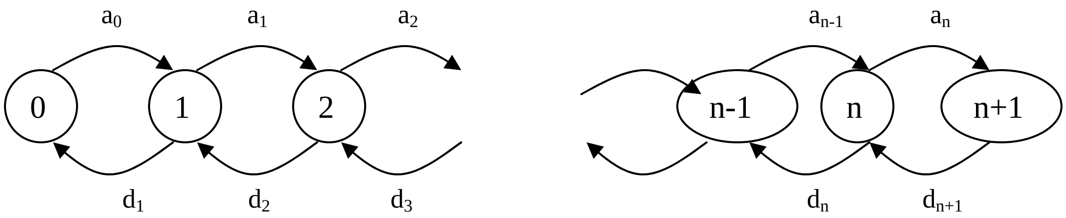
**3.** If the system is in state $S_n$ , at time t, the probability of a transition to $S_{n-1}$, in the time interval (t, t $+ \ \Delta$ t) is $d_n \ \Delta$ t. We refer to $d_n$ as the departure parameter (or the death parameter).

The process N(t) is sometimes referred to as the birth-death process.

Let $p_n(t)$ be the probability that the queueing system is in state $S_n$, at time t; that is,

$$p_n(t) = P\{N(t) = n\}$$

The state transition diagram for the birth-death process is shown in Fig. 16-2:



Where

$$p_1 = \frac{a_0}{d_1} p_0$$

$$p_2 = \frac{a_0 a_1}{d_1 d_2} p_0$$

$$p_n = \frac{a_0 a_1 .....a_{n-1}}{d_1 d_2 .......d_n} p_0$$

**THE M/M/1 QUEUEING SYSTEM**

In the M/M/1 queueing system, the arrival process is the Poisson process with rate $\lambda$ (the mean arrival rate) and the service time is exponentially distributed with parameter $\mu$ (the mean service rate).

Then the process N(t) describing the state of the M/M/1 queueing system at time t is a birth-death process with the Following state independent parameters:

$$a_n = \lambda \quad , \ n \geq 0 \quad , \ d_n = \mu \ , n \geq 1$$

Then

$$p_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho$$

$$p_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n = (1 - \rho)\rho^n$$

where $\rho = \frac{\lambda}{\mu} <$ 1, which implies that the server, on the average, must process the customers faster than their average arrival rate; otherwise the queue length (the number of customers waiting in the queue) tends to infinity. The ratio $\rho = \frac{\lambda}{\mu}$ is sometimes referred to as the traffic intensity of the system.

The average number of customers in the system is given by

$$L = \frac{\rho}{1 - \rho} = \frac{\lambda}{\mu - \lambda}$$

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

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{\rho}{\mu(1-\rho)}$$

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\rho^2}{1-\rho}$$

Examples:

1. Customers arrive at a watch repair shop according to a Poisson process at a rate of one per every 10 minutes, and the service time is an exponential r.v. with mean 8 minutes.
   ***(a)*** Find the average number of customers L, the average time a customer spends in the shop W, and the average time a customer spends in waiting for service $W_q$.
   *(b)* Suppose that the arrival rate of the customers increases 10 percent. Find the corresponding changes in L, W, and $W_q$.

(a) The watch repair shop service can be modeled as an **M/M/1** queueing system with $\lambda$ = 1/10 & $\mu$ = 1/8 . Thus, we have

$$L = \frac{\lambda}{\mu - \lambda} = \frac{1/10}{1/8 - 1/10} = 4$$

$$W = \frac{1}{\mu - \lambda} = \frac{1}{1/8 - 1/10} = 40 \quad \text{Minutes}$$

*$W_q$= W – $W_s$ = 40 - 8 = 32 minutes*

*(b) Now* $\lambda$ = 1/9 & $\mu$ = 1/8

$$L = \frac{\lambda}{\mu - \lambda} = \frac{1/9}{1/8 - 1/9} = 8$$

$$W = \frac{1}{\mu - \lambda} = \frac{1}{1/8 - 1/9} = 72$$

*$W_q$= W – $W_s$ = 72 - 8 = 64 minutes*

It can be seen that an increase of 10 percent in the customer arrival rate doubles the average number of customers in the system. The average time a customer spends in queue is also doubled.

**2. A** drive-in banking service is modeled as an **M/M/1** queueing system with customer arrival rate of 2 per minute. It is desired to have fewer than 5 customers line up 99 percent of the time. How fast should the service rate be?

P(5 or more customers in the system} = $\sum_{n=5}^{\infty} p_n = \sum_{n=5}^{\infty}(1-\rho)\rho^n = \rho^5 \qquad \rho = \frac{\lambda}{\mu}$

In order to have fewer than 5 customers line up 99 percent of the time, we require that this probability be less than 0.01. Thus,

$$\rho^5 = \left(\frac{\lambda}{\mu}\right)^5 \le 0.01$$

from which we obtain $\mu^5 \geq \frac{\lambda^5}{0.01} = \frac{2^5}{0.01} = 32000$

or $\mu \geq 5.024$

Thus, to meet the requirements, the average service rate must be at least 5.024 customers per minute.

3. People arrive at a telephone booth according to a Poisson process at an average rate of 12 per hour, and the average time for each call is an exponential r.v. with mean 2 minutes.
   (a) What is the probability that an arriving customer will find the telephone booth occupied?
   *(b)* It is the policy of the telephone company to install additional booths if customers wait an average of **3** or more minutes for the phone. Find the average arrival rate needed to justify a second booth.

(a) The telephone service can be modeled as an **M/M/1** queueing system with $\lambda$ = 1/5 & $\mu$ = 1/2

$\rho = \frac{\lambda}{\mu} = 2/5$.

The probability that an arriving customer will find the telephone occupied is P(L > 0), where L is the average number of customers in the system. Thus,

P(L > 0) = 1 – $p_0$ =1 – (1 - $\rho$ ) = $\rho$ =2/5 = 0.4

(b) $W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{\lambda}{0.5(0.5 - \lambda)} \geq 3$

from which we obtain $\lambda \geq$ 0.3 per minute. Thus, the required average arrival rate to justify the second booth is 18 per hour.