Queuing Theory

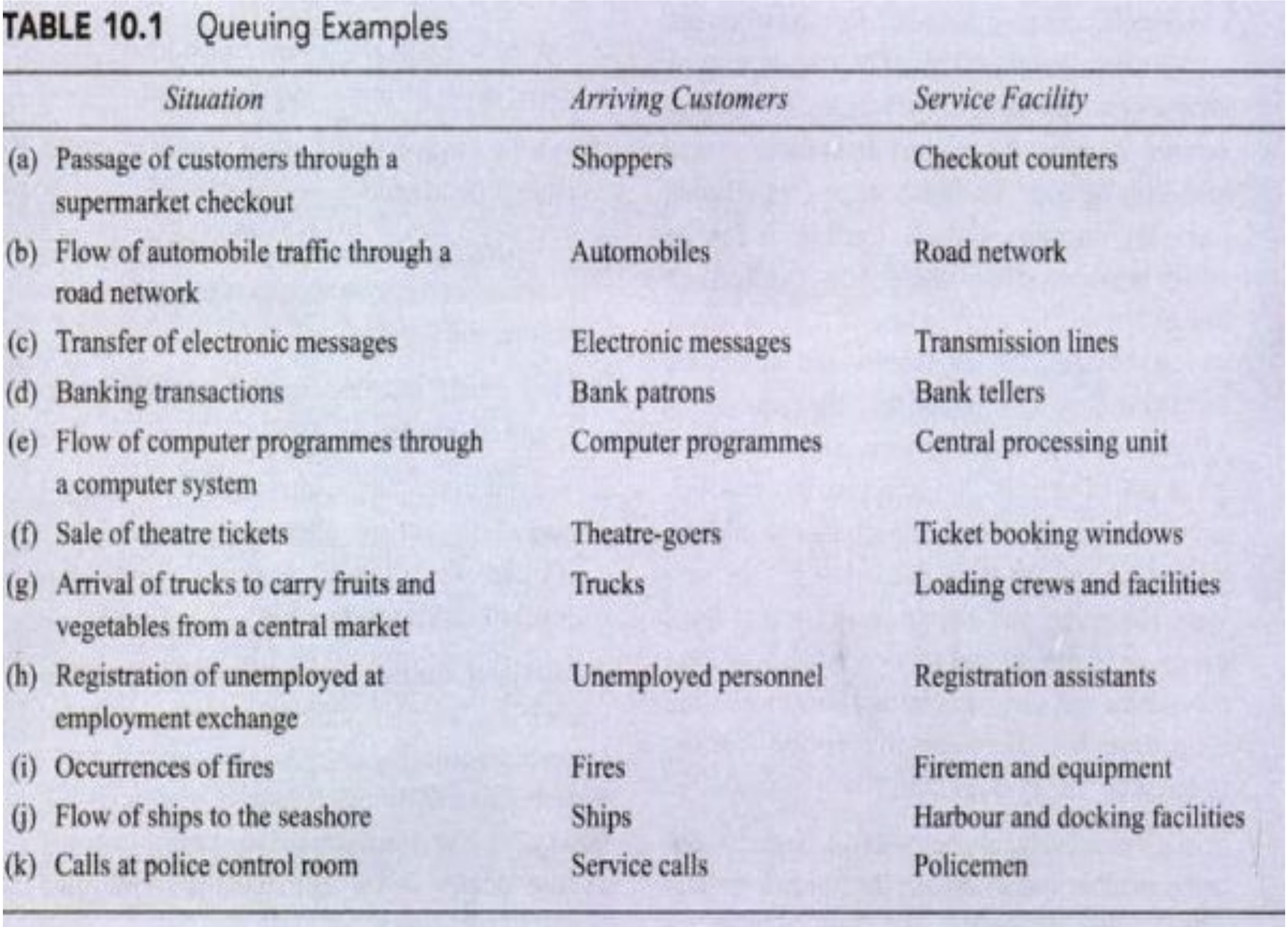

TABLE 10.1 Queuing Examples

Situation	*Arriving Customers*	*Service Facility*
(a) Passage of customers through a supermarket checkout	Shoppers	Checkout counters
(b) Flow of automobile traffic through a road network	Automobiles	Road network
(c) Transfer of electronic messages	Electronic messages	Transmission lines
(d) Banking transactions	Bank patrons	Bank tellers
(e) Flow of computer programmes through a computer system	Computer programmes	Central processing unit
(f) Sale of theatre tickets	Theatre-goers	Ticket booking windows
(g) Arrival of trucks to carry fruits and vegetables from a central market	Trucks	Loading crews and facilities
(h) Registration of unemployed at employment exchange	Unemployed personnel	Registration assistants
(i) Occurrences of fires	Fires	Firemen and equipment
(j) Flow of ships to the seashore	Ships	Harbour and docking facilities
(k) Calls at police control room	Service calls	Policemen

10.2 GENERAL STRUCTURE OF QUEUING SYSTEM

The general structure of a queuing system is depicted in Figure. 10.1.

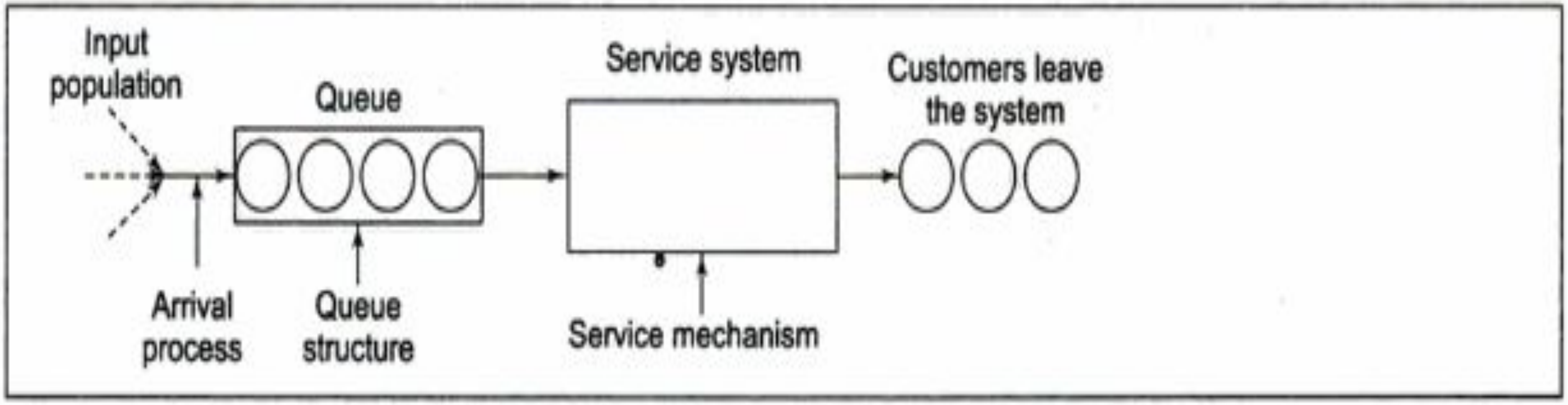

Figure 10.1 *General Structure of the Queuing System*

We shall discuss in more details the various elements of a queuing system and then present mathematical results for some specific systems. The elements of a system are:

1. Arrival Process The arrivals from the input population may be classified on different bases as follows:
(a) *According to source* The source of customers for a queuing system can be infinite or finite. For example, all people of a city or state (and others) could be the potential customers at a superbazar. The number of people being very large, it can be taken to be infinite. On the other hand, there are many situations in business and industrial conditions where we cannot consider the population to be infinite—it is *finite*. Thus, the ten machines in a factory requiring repairs and maintenance by the maintenance crew would examplify finite population. Removing one machine from a small, finite, population like this will have a noticeable effect on the calls expected to be made (for repairing) by the remaining machines than if there were a large number of machines, say 500.

(b) *According to numbers* The customers may arrive for service individually or in groups. Single arrivals are illustrated by customers visiting a beautician, students reaching at a library counter, and so on. On the other hand, families visiting restaurants, ship discharging cargo at a dock are examples of bulk, or batch, arrivals.

(c) *According to time* Customers may arrive in the system at known (regular or otherwise) times, or they might arrive in a random way. The queuing models wherein customers' arrival times are known with certainty are categorised as *deterministic models* (insofar as this characteristic is concerned) and are easier to handle. On the other hand, a substantial majority of the queuing models are based on the premise that the customers enter the system stochastically, at random points in time.

With random arrivals, the number of customers reaching the system per unit time might be described by a probability distribution. Although the arrivals might follow any pattern, the frequently employed assumption, which adequately supports many real world situations, is that the arrivals are *Poisson* distributed.

2. Service System There are two aspects of a service system—(a) structure of the service system, and (b) the speed of service.

(a) *Structure of the Service System*

By structure of the service system we mean how the service facilities exist. There are several possibilities. For example, there may be

(i) A Single Service Facility A library counter is an example of this. The models that involve a single service facility are called *single server models*. Figure 10.2 (a) illustrates such a model.

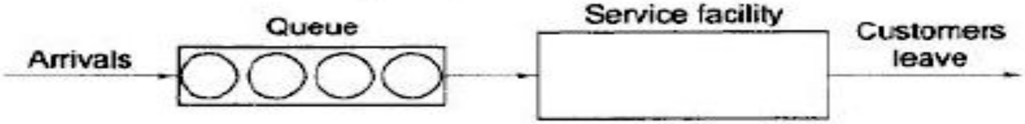

Fig. 10.2(a) Single Server, Single Queue Model

(ii) Multiple, Parallel Facilities with Single Queue That is, there is more than one server. The term parallel implies that each server provides the same type of facility. Booking at a service station that has several mechanics, each handling one vehicle, illustrates this type of model. It is shown in Fig. 10.2 (b).

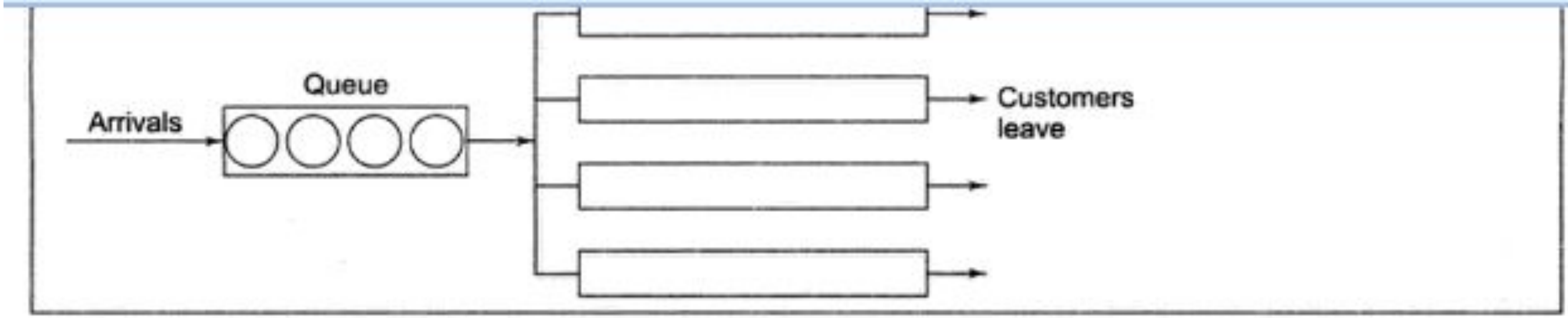

Figure 10.2(b) *Multiple, Parallel Servers, Single Queue Model*

(iii) Multiple, Parallel Facilities with Multiple Queues This type of model is different from the earlier one only in that each of the servers has a different queue. Different cash counters in an electricity office where the customers can make payment in respect of their electricity bills provide an example of this type of model. Figure 10.2(c) portrays such a model.

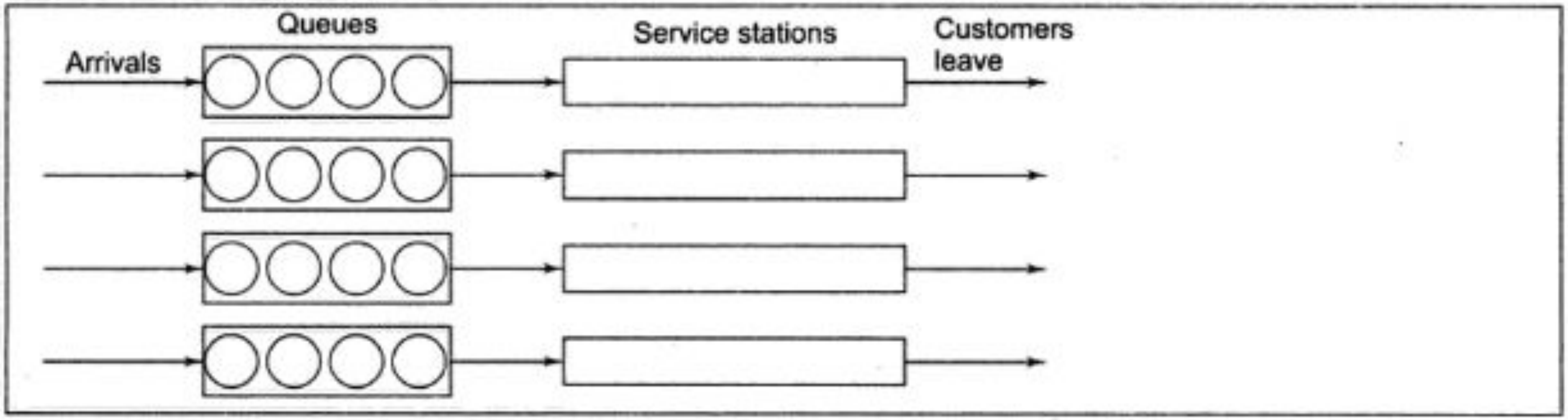

Figure 10.2(c) *Multiple, Parallel Servers, Multiple Queues Model*

(iv) Service Facilities in a Series In this, a customer enters the first station and gets a portion of service and then moves on to the next station, gets some service and then again moves on to the next station ... and so on, and finally leaves the system, having received the complete service. For example, machining of a certain steel item may consist of cutting, turning, knurling, drilling, grinding, and packaging operations, each of which is performed by a single server in a series. Figure 10.2(d) shows such a situation.

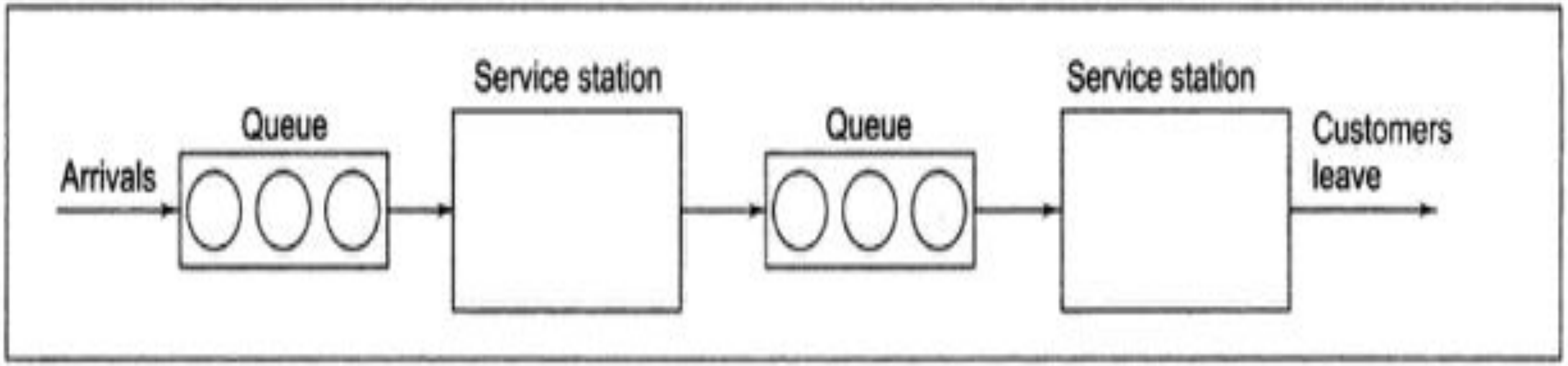

Figure 10.2(d) *Multiple Servers in Series*

Besides these, there may be other possibilities as well.

(b) Speed of Service

In a queuing system, the speed with which service is provided can be expressed in either of two ways—as *service* rate and as *service time*. The service rate describes the number of customers serviced during a particular time period. The service time indicates the amount of time needed to service a customer. Service rates and times are reciprocals of each other and either of them is sufficient to indicate the capacity of the facility. Thus, if a cashier can attend, on the average, to 10 customers in an hour, the service rate would be expressed as 10 customers/hour and service time would be equal to 6 minutes/customer. Generally, however, we consider the service time only.

If these service times are known exactly, the problem can be handled easily. But, as generally happens, if these are different and not known with certainty, we have to consider the distribution of the service times in order to analyse the queuing system. Generally, the queuing models are based on the assumption that service times are *exponentially* distributed about some average service time.

3. Queue Structure Another element of a queuing system is the queue structure. In the queue structure, the important thing to know is the queue discipline which means the order by which customers are picked up from the waiting line for service. There are a number of possibilities. They are:

(a) First-come-first-served

When the order of service of customers is in the order of their arrival, the queue discipline is of the first-come-first-served type. For example, with a queue at the bus stop, the people who came first will board the bus first.

(b) Last-come-first-served

Sometimes, the customers are serviced in an order reverse of the order in which they enter so that the ones who join the last are served first. For example, assume that letters to be typed, or order forms to be processed accumulate in a pile, each new addition being put on the top of them. The typist or the clerk might process these letters or orders by taking each new task from the top of the pile. Thus, a just arriving task would be the next to be serviced provided that no fresh task arrives before it is picked up. Similarly, the people who join an

(c) Service-in-random-order (SIRO)

Random order of service is defined as: whenever a customer is chosen for service, the selection is made in a way that every customer in the queue is equally likely to be selected. The time of arrival of the customers is, therefore, of no consequence in such a case.

(d) Priority Service

The customers in a queue might be rendered service on a priority basis. Thus, customers may be called according to some identifiable characteristic (length of job, for example) for service. Treatment of VIPs in preference to other patients in a hospital is an example in point.

For the queuing models that we shall consider, the assumption would be that the customers are serviced on the first-come-first-served basis.

Another thing to consider in the queuing structure is the behaviour or attitude of the customers entering the queuing system. On this basis, the customers may be classified as being (a) patient, or (b) impatient. If the customers join a queue, when it exists, and wait till they enter the service station for getting service, they are called *patient* customers. On the other hand, the queuing systems may enjoy customer behaviour in the form of defections from the queue. The customers may not select queues randomly (if there are multiple queues) and look for the shortest queue. There may be *jockeying* among the many queues, that is the customers may switch to other queues which are moving 'fast', and also *reneging* is possible—when a customer stands in the queue for sometime and then leaves the system because it is working 'too slowly'. There may also be *bribing* or *cheating* by some customers for queue positions. Besides, some customers may, upon their arrival, not join the queue for some reason and decide to return for service at a later time, or may even abandon the input population altogether. In terms of the queuing theory, this is known as *balking*, and occurs particularly when there are limits on the time and the extent of storage capacity available to hold waiting customers. Unless otherwise specified, the storage capacity is taken to be infinite. In the queuing models that we consider, we shall assume that there is no balking or jockeying and that the customers leave the system only after receiving service, and not before. Mathematical models give way to simulation when this assumption breaks.

10.3 OPERATING CHARACTERISTICS OF QUEUING SYSTEM

An analysis of a given queuing system involves a study of its different operating characteristics. This is done using queuing models. Some of the more commonly considered characteristics are discussed below.

1. *Queue length*—the average number of customers in the queue waiting to get service. Large queues may indicate poor server performance while small queues may imply too much server capacity.

2. *System length*—the average number of customers in the system, those waiting to be and those being serviced. Large values of this statistic imply congestion and possible customer dissatisfaction and a potential need for greater service capacity.

3. *Waiting time in the queue*—the average time that a customer has to wait in the queue to get service. Long waiting times are directly related to customer dissatisfaction and potential loss of future revenues, while very small waiting times may indicate too much service capacity.

4. *Total time in the system*—the average time that a customer spends in the system, from entry in the queue to completion of service. Large values of this statistic are indicative of the need to make adjustment in the capacity.

5. *Server idle time*—the relative frequency with which the service system is idle. Idle time is directly related to cost. However, reducing idle time may have adverse effects on the other characteristics mentioned above.

We now proceed to discuss some of the queuing models. It may be mentioned here that the results obtained from various models are based on the assumption that the service system is operating under equilibrium or *steady state* conditions. For many systems, the operating day begins in *transient state* with no customers in the system. It takes some initial time interval for enough customers to arrive such that a steady state balance is reached. It should be clearly understood that a steady state does not mean that the system will reach a point where the number of customers in the system never changes. Even when the system reaches equilibrium, fluctuations will occur. A steady state condition really implies that various system performance measures (the operating characteristics) would reach stable values.