SIMULATION OF QUEUING SYSTEMS

6.1 Introduction

Waiting line queues are one of the most important areas, where the technique of simulation has been extensively employed. The waiting lines or queues are a common site in real life. People at railway ticket window, vehicles at a petrol pump or at a traffic signal, workers at a tool crib, products at a machining center, television sets at a repair shop and consumers at a ration depot are a few examples of waiting lines. The waiting line situations arise, either because,

- There is too much demand on the service facility so that the customers or entities have to wait for getting service, or
- There is too less demand, in which case the service facility have to wait for the entities.

Thus, when facilities are inadequate, the entities wait and incur cost due to waiting time, and when the demand is inadequate, the facilities wait and incur cost due to idle time. The objective in the analysis of queuing situations is to balance the waiting time and idle time, so as to keep the total cost at minimum.

The queuing theory owe its development to an engineer named A.K. Erlang, who in 1920, studied waiting line queues of telephone calls in Copenhagen, Denmark. The problem was that during the busy period, telephone operators were unable to handle the calls, there was too much waiting time, which resulted in consumer dissatisfaction.

Many researchers carried on the research work in the telephone traffic further and it was only after World War II, that the queuing theory encompassed the waiting line situations from other fields as business and industry.

6.2 The Components of a Waiting Line System

(*i*) **Cálling Source** or the population from which customers are drawn. The calling source may be finite or infinite. When the arrival of a customer does not affect the next arrival, the source of customers is said to be infinite. Population of vehicles arriving at a toll booth or patients arriving at a hospital have infinite calling source. In case of a repair crew looking after a group of 5 machines, if one breaks down and put under repair, then the next breakdown has to come from only 4 machines, the probability of which will be more than the previous. The population in this case is finite.

(*ii*) **Waiting Line** or queue, *i.e.*, the number of customers waiting to be served. Depending upon the space available, it may again be of finite or infinite length. In a barbershop having four chairs, the queue can have a maximum length of four, because people seldom queue up outside a barber shop. An important case of finite queue is the buffer used to decouple two sequential workstations, where the output of first station is not perfectly matched to the second workstation.

(*iii*) **Service Facility** or the number of service channels. The simplest case is of one service channel, where all the customers form one queue and are attended by one server, a single server model. In many cases, waiting line systems have more than one service facility.

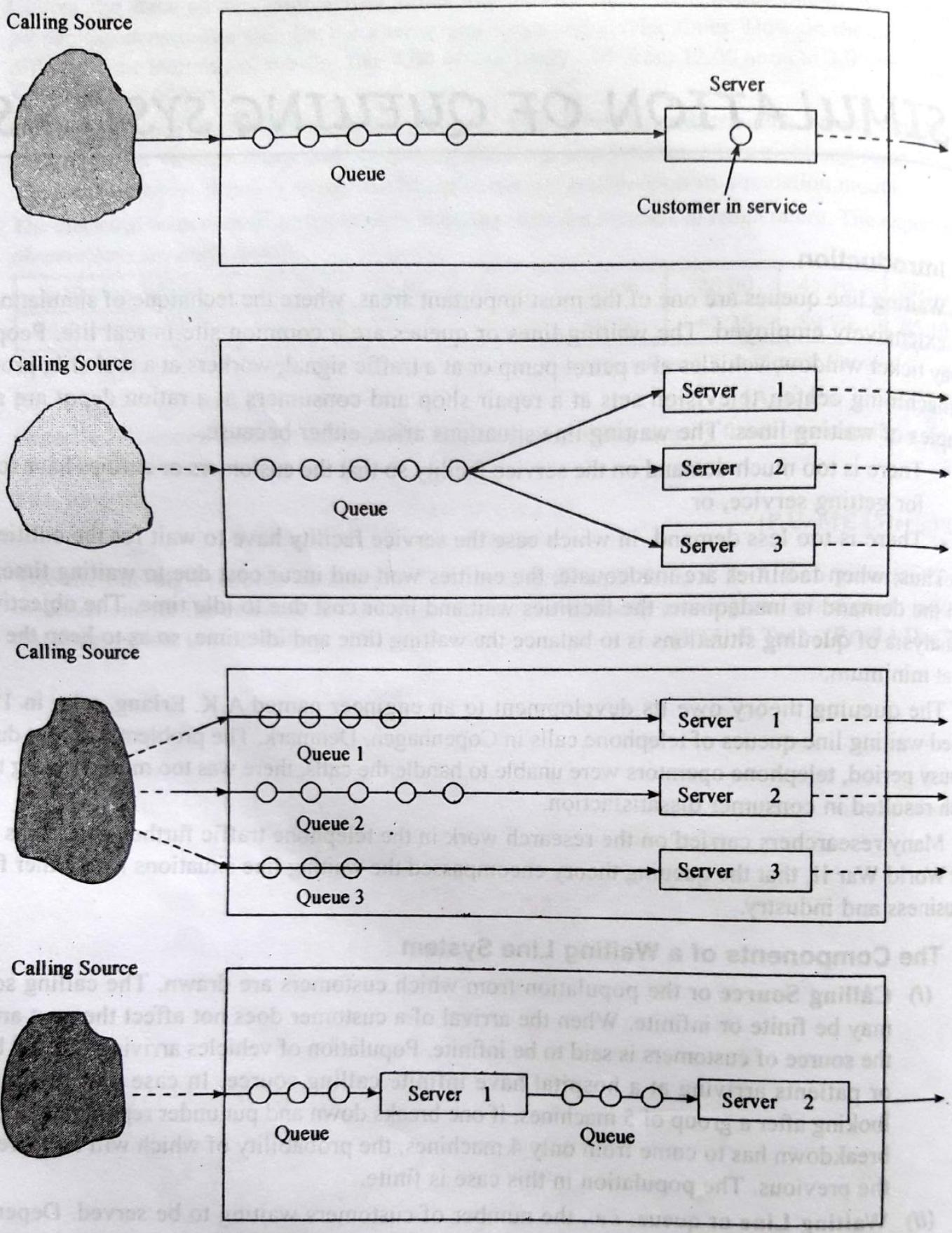

Fig. 6.1

These facilities may be working in parallel or in series. When in parallel, there may be a single queue or multiple queues. Some typical queuing systems are shown in Fig. 6.1.

The important **attributes** that determine the properties of a waiting line queue are,

- Input or arrival rate.
- Output or service rate.
- Service or queue discipline.

The arrival rate is the average number of entities, which join the waiting line per unit time. Depending upon the system, the time unit may be a second, a minute, an hour or a day etc. The average time between consecutive arrivals is called **inter-arrival time.** In most of the situations, the arrivals are a random phenomenon. Different probability density functions are applicable to different arrival patterns, but a commonly made assumption is that arrival rates follow the Poisson distribution.

The service rate or the departure rate is the average number of customers served per unit time. Average service time is the reciprocal of the service rate. Depending upon the situation, the service rate may be constant or variable and may follow any distribution as Uniform, Normal, Exponential, and Erlang etc. But in most of the situations service time is assumed to follow exponential distribution.

The third factor for describing the waiting line is the **queue discipline** that determines how the customers are selected from the queue for service. The most commoh queue disciplines are given below.

(*i*) **First-in, First-out (FIFO):** Also called the first come first served, is the most common service discipline, according to which the customers are served in the order of their arrival.

(*ii*) **Last-in, First-out (LIFO):** In some situations, the last arrival is served first, as in big go-downs the items coming last are taken out first, in crowded trains or elevators, passengers getting in last come out first.

(*iii*) **Priority:** An arrival may be given priority over the customers waiting in line. A particular machine in a production shop may be more important than the others, and when it breaks down, its repair may be taken up on priority as compared to the other broken down machines. When an arrival, not only goes to the head of the queue, but displaces any unit already in service, it is said to have **pre-emptive priority**. The new arrival is said to pre-empt or interrupt the service.

(*iv*) **Random:** The service discipline is said to be random, when all waiting customers have equal chance of getting selected for service. The selection follows purely random choice.

Some other terms commonly associated with waiting lines are given below.

- **Reneging:** When a queue grows excessively long, a customer waiting in the queue may become impatient and may leave the queue before it is due to enter the service facility. This process is called reneging.
- **Balking:** When a queue grows very long and an arrival refuses to join the queue, it is called balking.
- **Jockeying:** In multiple queues before multiple service channels, where all the channels are providing the same service, a customer may leave one queue and join the other looking faster. This process is called jockeying.
- **Polling:** When there are more than one queues forming for the same service, the action of sharing service between the queues is called polling. A bus picking up passengers from different stoppages along its route is an example of polling service. Separate queues for ladies and gents at a ticket window, is another example of polling service.

6.5 Measures of System Performance

The performance of a queuing system can be evaluated in terms of a number of response parameters, however the following four are generally employed.

(*i*) Average number of customers in the queue or in the system.

(*ii*) Average waiting time of the customers in the queue or in the system.

(*iii*) System utilization.

(*iv*) The cost of waiting time and idle time.

Each of these measures has its own importance. The knowledge of average number of customers in the queue or in the system helps to determine the space requirements of the waiting entities. Also too long a waiting line may discourage the prospectus customers, while no queue may suggest that service offered is not of good quality to attract customers.

The knowledge of average waiting time in the queue is necessary for determining the cost of waiting in the queue.

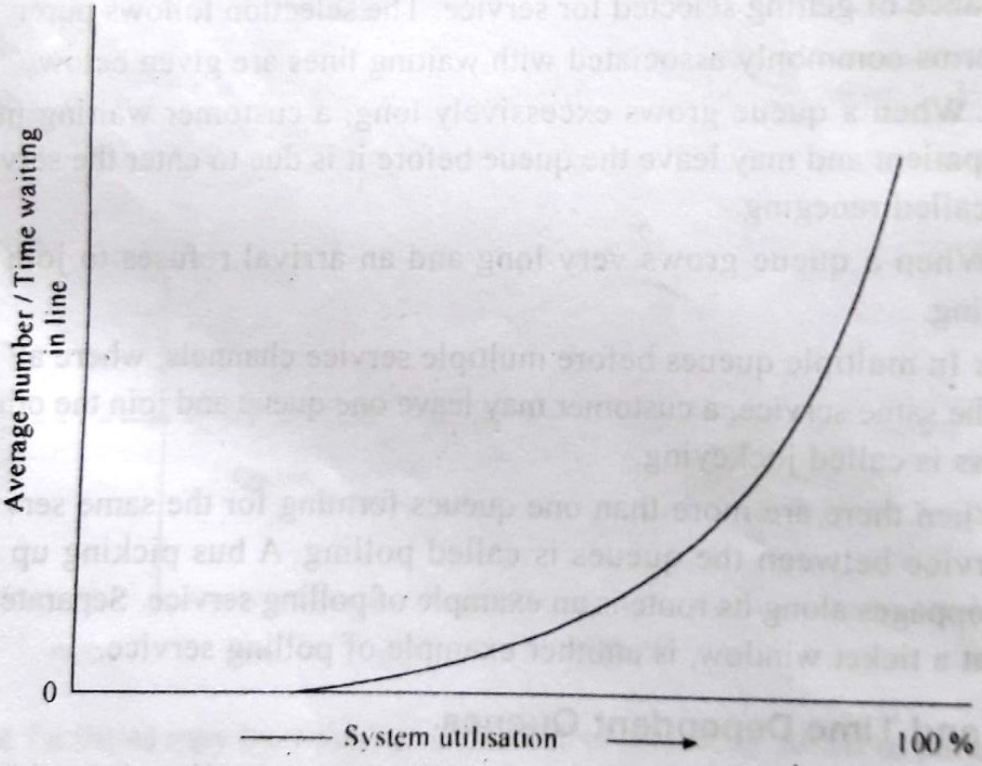

Fig. 6.2

System utilization, that is, the percentage capacity utilized reflects the extent to which the facility is busy rather than idle. System utilization factor (S) is the ratio of average arrival rate (λ) to the average service rate (μ).

$S = \lambda / \mu$ in case of a single server model.

$= \lambda / \mu n$ in case of a 'n' server model.

The system utilization can be increased by increasing the arrival rate which amounts to increasing the average queue length as well as the average waiting time, as shown in Fig. 6.2. Under normal circumstances 100% system utilization is not a realistic goal.

6.6 Costs of Customer

6.7 Kendall's Notation

Kendall's Notation for specifying the characteristics of a queue is V/W/X/Y/Z, where,

V indicates the arrival pattern,
W indicates the service pattern,
X gives the number of servers,
Y represents the system capacity, and
Z indicates the queue discipline.

The symbols used for the inter-arrival times, service times, and the queue discipline are,

Queue characteristic	*Symbol*	*Meaning*
Inter-arrival time or Service time	D	Deterministic
	M	Exponential
	Ek	Erlang with value of parameter k
		Any other distribution
Queue discipline	FIFO	First-in First-out
	LIFO	Last-in First-out
	SIRO	Service in random order
	PRI	Priority ordering
	GD	Any other specified ordering

If the capacity Y is not specified, it is taken to be ∞ (infinite) and if the queue discipline is not specified, it is taken as FIFO.

An M/D/2/5/FIFO stands for a queuing system having exponential arrival times, deterministic service times, two servers, with a capacity of 5 customers and having the first in first out queue discipline.

And an M/D/2 will mean exponential inter-arrival times, deterministic service times, two servers, infinite system capacity and FIFO queue discipline.