Lecture 18
Chapter # 6
FUNDAMENTALS OF QUEUING MODELS

What is Queuing Theory?

- Queuing theory is the study of waiting in various guises.

- It deals with quantifying the phenomenon of waiting in lines using representative measures of performance, such as
  - average queue length,
  - average waiting time in queue, and
  - average service time

- It uses queuing models to represent the various types of queuing systems that arise in practice.

And these models enables us finding an appropriate balance between the cost of service and the amount of waiting by the customer.

We'll be learning about various formula for every model & these formulas actually indicates how the corresponding queuing system should perform.

- *Advantages of queuing models:* very helpful for determining how to operate a queuing system in the most effective way.
- Providing too much service capacity to operate the system involves excessive costs. The overall cost of the system will tremendously increase.
- But not providing enough service capacity results in excessive waiting and all its unfortunate consequences.

Here the key point is to keep a balance between the waiting time by the customers and the average service provided by the system.

- The operating characteristics of queuing systems are determined largely by two statistical properties, namely,
  - the probability distribution of *inter-arrival* times and
  - the probability distribution of *service* times.

The arrival of customers or the arrival of people in a queue are totally random. Sometimes they could be deterministic as well but it's a rare case. In practical situation or in a normal scenario, the arrival of customers in a queue is totally probabilistic or totally random. Therefore, there inter-arrival times are also random.

Service time; i.e. how much service is allocated to be every customer.

The Basic Queuing Process

- *Customers* requiring service are generated over time by an *input source*.
- These customers enter the *queuing system* and join a *queue*.
- At certain times, a member of the queue is selected for service by some rule known as the *queue discipline*.

- The required service is then performed for the customer by the *service mechanism* after which the customer leaves the queuing system.

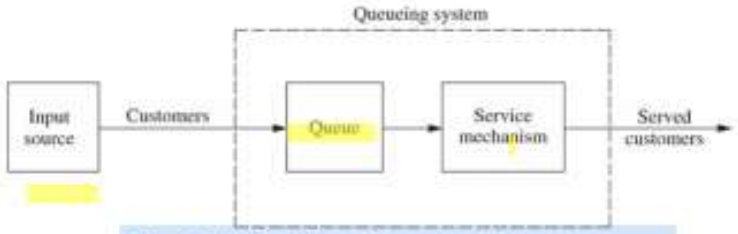


Fig 1: The basic queuing process

Input Source (Calling Population)

- Calling population: population from which arrivals come.
- One characteristic of the input source is its size.

o total number of distinct potential customers that might require service from time to time.

- The size may be assumed to be either *infinite* or *finite*.

Infect in the calculation or solution of numerical, the assumptions of infinite population size are quite useful.

- An infinite source is forever abundant (e.g., calls arriving at a telephone exchange).
- Because the calculations are ***far easier*** for the infinite case, this assumption often is made even when the actual size is some relatively large finite number.
- The finite case is ***more difficult analytically.***

Because the number of customers in the queueing system at any point in time actually effects the total number of potential customers that are outside the system.

- The statistical pattern by which customers are generated over time must also be specified.
- The common assumption: *Poisson process*; i.e.,

- the number of customers generated until any specific time has a Poisson distribution.

- arrivals to the queuing system occur randomly but at a certain fixed mean rate, regardless of how many customers already are there
- so the *size* of the input source is *infinite*.

An equivalent assumption: the probability distribution of the time between consecutive arrivals is an *exponential distribution.*
The Poisson processes and the exponential distributions are actually used to model the queueing system.

Queuing Behavior

- The queuing behavior of customers plays a role in waiting-line analysis.
- Human customers may ***jockey*** from one queue to another in the hope of reducing waiting time.

Jockey means rider. He's a person who usually rides in horse races. So sometimes what we do, we try to ride from one queue to another so that the waiting time that we have already anticipated could be reduced.

- They may also **balk** from joining a queue altogether because of anticipated long delay, or

Balks means hesitant or to be unwilling to accept an idea. it means that sometime people may be hesitant to join a queue at the first place because they have already foreseen the long waiting times and long delay.

- They may **renege** from a queue because they have been waiting too long.

Renege means they can go back on something. Let's suppose I am standing in a queue and it's been a long time waiting for the service so I may go back and decided to eventually come back later for the service.

Queue

- The queue is where customers wait *before* being served.
- A queue is characterized by the maximum permissible number of customers that it can contain.
- Queues are called *infinite* or *finite*, according to whether this number is infinite or finite.
- The assumption of an *infinite queue* is the standard one for most queuing models.

▪ However, for queuing systems where this upper bound is small enough that it actually would be reached with some frequency, it becomes necessary to assume a *finite queue.*

Queue Discipline

The order in which members are selected from a queue

▪ An important factor in the analysis of queuing models.

▪ First-come-first-served (FCFS) usually is assumed by queuing models, unless stated otherwise.

▪ Other disciplines include Last Come, First Served (LCFS) and Service In Random Order (SIRO).

▪ Customers may also be selected from the queue based on some order of priority.

▪ For example, rush jobs at a shop are processed ahead of regular jobs.

Or in bank old people are preferred over young people.

Service Mechanism

▪ The service mechanism consists of

- one or more *service facilities*,
- each of which contains one or more *parallel service channels*, called *servers*.

Sometimes in the service mechanism there could be just a single server & usually it consists of two or more parallel service channels to expedite the process.

▪ If there is more than one service facility, the customer may receive service from a sequence of these (*service channels in series*).

Sometimes the service channels are in series & customers may receive some of the services from single channel & the remaining service from the other channel and so on.

▪ At a given facility, the customer enters one of the parallel service channels and is completely serviced by that server.

Sometimes it is also possible that the service channels are purely independent & are running parallel once the customer enters a certain service channel it will eventually leave the system after getting all the required services.

▪ A queuing model must specify the arrangement of the facilities and the number of servers (parallel channels) at each one.

▪ Most elementary models assume one service facility with either one server or a finite number of servers.

▪ **Service time (or holding time)**: The time elapsed from the commencement of service to its completion for a customer at a service facility.

▪ The service-time distribution that is most frequently assumed in practice

- (largely because it is far more tractable than any other)
- is the *exponential* distribution, and most of our models will be of this type.

▪ Other important service-time distributions are

- the *degenerate* distribution (constant service time) and
- the *Erlang (gamma)* distribution.

Cost-based Queuing Decision Model

➢**Cost optimization model:** we seek the minimization of the sum of the two costs:-

- the cost of offering the service and the
- cost of waiting.

▪ Fig 2 depicts a typical cost model (in Rs per unit time).

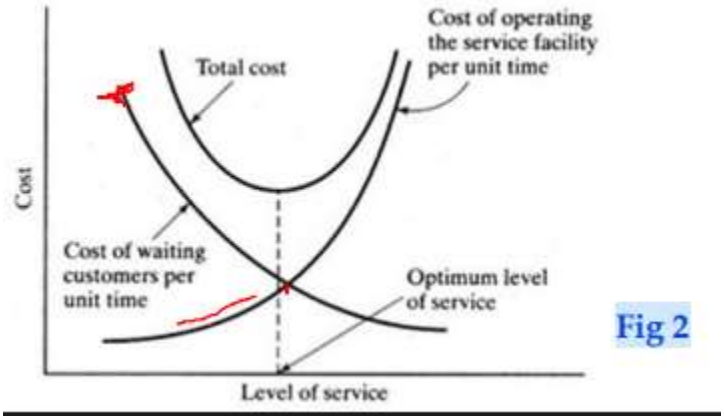


Fig 2

Initially when the customers it actually starts from here the cost is being represented by y-axis. & the level of customers using x-axis. So initially when the customers are waiting in the queue the cost of waiting time is very high. The cost of service is low at that particular point in time when the service is just starting its facility. Once the number of customers are being provided with the service the level of service increases & the cost of waiting is decreasing. This intersection point is actually representing the optimal level of service & this combined curve or this distribution is representing the total cost of the service.

The total cost is actually the sum of the Cost of waiting customers plus the cost of operating the service facility.

▪ **The main obstacle:** difficulty of obtainin waiting, particularly when human behavior is involved.

An Elementary Queuing Process

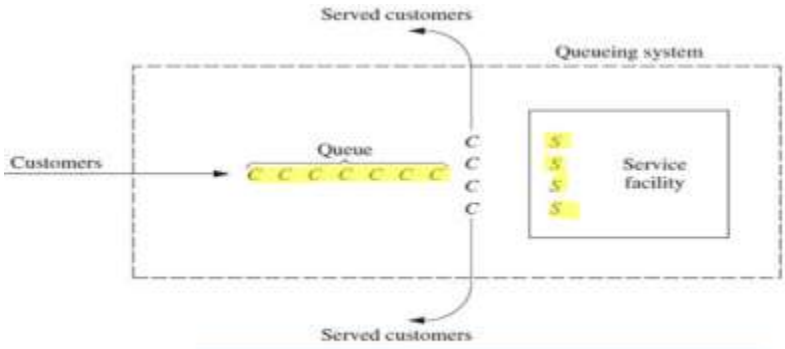
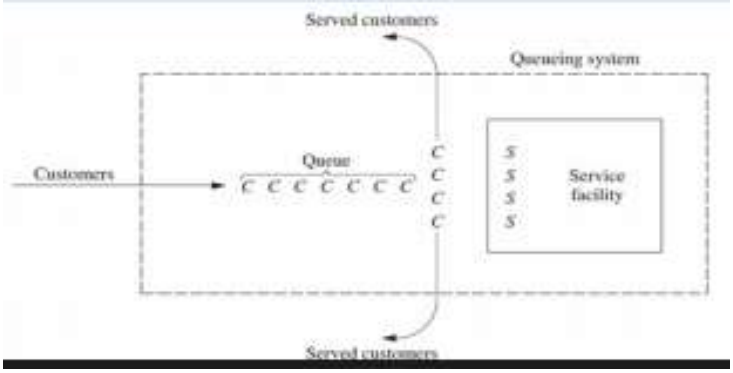


Fig 3: An elementary Queuing System

- Queuing theory has been applied to many different types of waiting-line situations.
- The most prevalent type of situation is the following:

- A single waiting line (which may be empty at times) forms in the front of a single service facility, within which are stationed one or more servers.
- Each customer generated by an input source is serviced by one of the servers, perhaps after some waiting in the queue (waiting line).
  - Furthermore, servers need not even be people.
  - In many cases, a server can instead be a machine, a vehicle, an electronic device, etc.
  - By the same token, the customers in the waiting line need not be people.
  - For example, they may be items waiting for a certain operation by a given type of machine, or they may be cars waiting in front of a tollbooth.



Moreover, you may also learn in different literatures that it is not necessary that there actually be a physical waiting line forming in front of a physical structures that constitutes the service facility. The members of the queue may instead be scattered throughout in area waiting for server to come to them. For example, there are various machines waiting to be repaired by the human here the customer that actually needs service is the machine the car that is need to be repaired by some human. The server or group of servers that are actually assigned to a given area constitutes the service facility for that area.

Still queueing theory will give us the average number of customers who are waiting, the average waiting time & so on.
The only essential requirement for queueing theory to be applicable is that the changes in the number of customers waiting for a given service occurs just as those physical situations that is being described in the given figure.
This was all about the elementary queueing process infect we have already learned about the basic queueing structure in our OS course where we have seen that the service facilities being represented by the processor or may be represented by the IO channels and the customers are the processes that are waiting in a queue & by means of a certain queueing discipline they are provided the service facility by a processor.

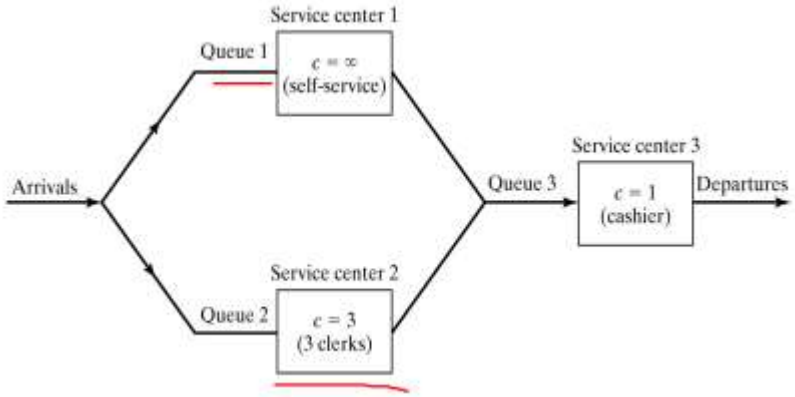
Examples

System	Customers	Server
Reception desk	People	Receptionist
Hospital	Patients	Nurses
Airport	Airplanes	Runway
Production line	Cases	Case-packer
Road network	Cars	Traffic light
Grocery	Shoppers	Checkout station
Computer	Jobs	CPU, disk, CD
Network	Packets	Router

Example Problem

Consider a discount warehouse where customers may:
• serve themselves before paying at the cashier (service center 1) or
• served by a clerk (service center 2)
Solution:
Here you need to draw a queueing system.
3rd service center could be the cashier.
C= infinite; we are actually observing the self-service phenomena where any customer can come and serve himself.
C=3 ; service capacity is 3 so there are 3 clerks available to serve the customers.

After completing the required service, the customers need to pay as well for that pay customers may form a queue so there is a cashier for their assistance & there's just one cashier in that particular ware house. So c=1
After complete operation the customers may depart.



Here I have replaced the service center 2 by three parallel servers; it means that at any one point in time the customer may be served by either 1st clerk or 2nd clerk or 3rd clerk.