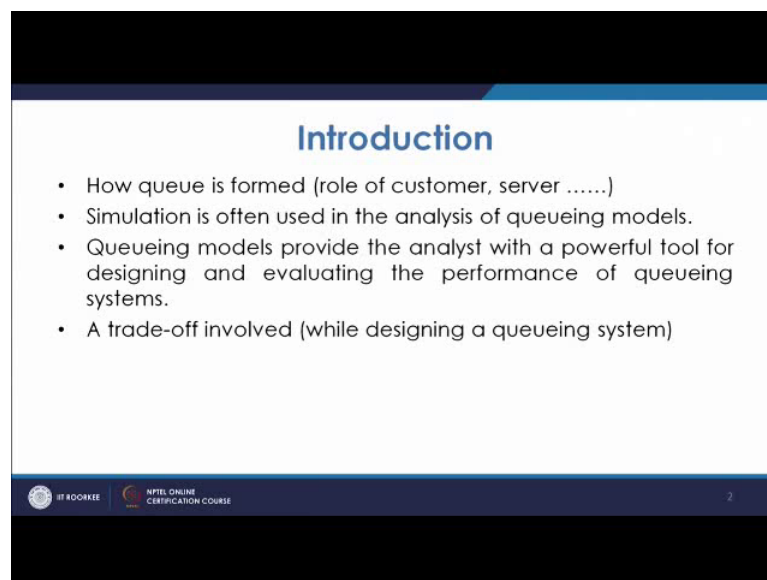


**Modeling & Simulation of Discrete Event Systems**  
**Dr. Pradeep K Jha**  
**Department of Mechanical and Industrial Engineering**  
**Indian Institute of Technology, Roorkee**

**Lecture – 11**  
**Characteristics of a Queueing System**

Welcome to the lecture on characteristics of a queueing system. So, this course modeling and simulation of discrete event system. So, queueing is one of the examples of the events, where the events take place in discrete manner and we will have some introduction about the queueing system, what a queue is? What are the main characteristics of the queue and how it is formed?

(Refer Slide Time: 00:57)



**Introduction**

- How queue is formed (role of customer, server .....)
- Simulation is often used in the analysis of queueing models.
- Queueing models provide the analyst with a powerful tool for designing and evaluating the performance of queueing systems.
- A trade-off involved (while designing a queueing system)

IT ROORKEE    NPTEL ONLINE CERTIFICATION COURSE    2

So, we will discuss in some detail in this lecture. So, the first thing is how a queue is formed. So, basically, we come across a queue in our daily life, we go to any counter and we are in the queue, if there are persons already standing, there and if there are no person is standing, it means there is no queue. So, you are going directly to get the service. So, basically, for formation of a queue, you need the customer. The customer can be anyone, he may be a person, he may be a machine, he may be any entity, which has to be served and then there has to be a server. Server means the one who is giving the service.

So, he is at the point, from here he will be giving the service. So, there has to be a customer and there has to be a server and then queue has a meaning; no. Why queue

analysis is required. So, basically what we see that in our daily life, because you know the organization needs to give the service and from every service, it has to earn people, pay for the service, they get now the thing is that, if the server is always free, means the organization has to pay fee.

So, that in organization cannot, you know effort to do. So, the thing is that, since people need service, people are in need of service. They have to go at the service counter. So, that is how a server is important. Now, depending upon the importance, you will have to spend time in the queue. So, if you feel that, this work is very much important. You would like to spend more time in the queue, if you feel that the worth of spending time in the queue is not. So, high then the importance of the, you know that you want in that case, you want like to go to queue.

So, you will see that without that, your work is done. Now, there are certainly you know principles, if there are small queues, if there is queue and you have the details for suppose, 5 minutes, 10 minutes and there is simple arithmetic involved, you can do the analysis, but then in today's life this queue has become very important, you see that large queues are there, if there is competition in the market, people try to see that the customers should not wait for long.

So, they will have to optimize. So, for that and then there are many things, many parameters are involved which are basically not deterministic, they are probabilistic. So, many a times, there is complex mathematics involve complexities are there. So, this analysis requires, the use of powerful computing devices and that is how when whenever, we use this computing devices. We do the computations, we do the simulations, that is, what is the simulation is. So, whenever you do. So, as we will see if you have studied the operations research course or anything about queueing you must be hearing or you must be knowing about different type of queues.

So, in that basically, there are lot of parameters, there are lot of conditions and they need to be solve. So, as long it is simple, you can do it using normal mathematics, but then if it is complex, then you will have to go for simulation. So, queueing models provide the analyst with a powerful tool for designing and evaluating the performance of queueing system. So, you have different kind of models and these models help you to understand

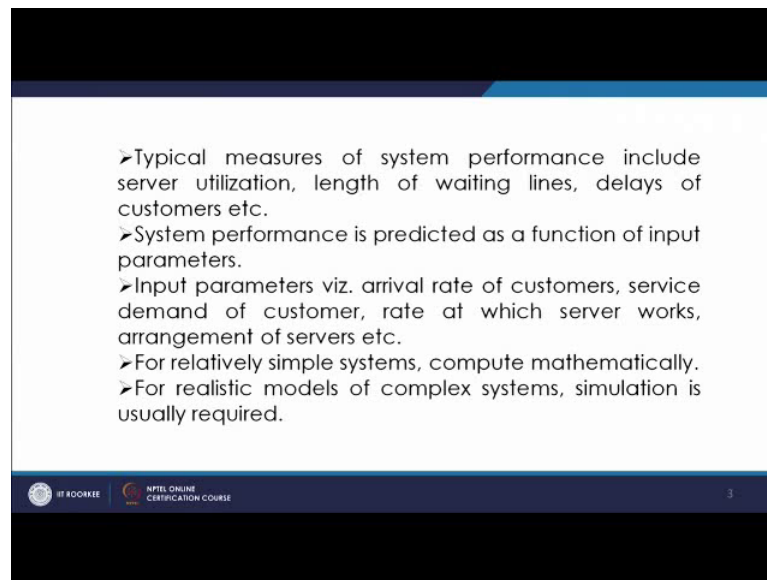
and evaluate the performance of a queueing system. There are many ways by which a system may be evaluated in queue, like how much time a customer has to wait.

On an average, how much customer how many customers are coming and they are getting the service or on an average, how much one customer spends time in the system or in the queue all these are the different you know performance measures. So, that we will discuss later and basically, the models which we make that is a tool for the analyst. So, that he can work on it he can think about it and he can come to the conclusions, you, he can come with certain suggestions. Now, a trade off is involved, while designing a queueing system, what is that trade off. So, as we discussed that when a queue is formed there are the 2 persons; one is customer and other is server. The customer feels always that whenever you should go, it should be, you know minimum time for which he should wait.

So, ideally he feels that whenever you goes into the queue, he should simply get the service means, there should no one ahead of him he should get the service whereas, the server, who is sitting the server or the organization. They feel that for the whole duration, there should be large number of queues. So, that they see that, there is completely utilization of the server, there is complete utilization of the machine. So, that they can earn maximum. They can give more and more number of service and they can get large earnings. So, basically both are the extreme ends and you will have to have a trade off. You will have to have a balance, the customer will not mind, if he feels that he has to wait for, suppose, 5 minutes.

If he has to wait for 5 minutes or if you, has to go for more than more than that then he may leave. So, that will be a lost to the organization or the server. So, and the server will also see. So, and in that case customer we will see that I can wait for some time. So, customer has to compromise little bit whereas, the server will also see that the queue, the person should not wait for more than certain time. So, that way there is a tradeoff and then a queueing system is designed, so that both party, the customer as well as the server or the organization, I mean aspect of this organization, itself they do not suffer to the extreme extent.

(Refer Slide Time: 08:24).



- Typical measures of system performance include server utilization, length of waiting lines, delays of customers etc.
- System performance is predicted as a function of input parameters.
- Input parameters viz. arrival rate of customers, service demand of customer, rate at which server works, arrangement of servers etc.
- For relatively simple systems, compute mathematically.
- For realistic models of complex systems, simulation is usually required.

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 3

So, typical measures of system performance. So, what are the normal I mean typical system performance measures? So, they include the server utilization. Server utilization means for what fraction of the time, the server is utilized. So, this is one of the typical measures of system performance or output measures of performance are the ones for which we are interested to find the value. So, in the queueing system, normally we find the server utilization, we find length of waiting lines, delays of customers, average number of customers in the queue, all that average number of customers in the system.

These are the typical output performance measures in a queueing system. System performance is predicted as a function of input parameters, means you know we have to express the system performance and you have certain input parameters, some of the things you know, input parameters means those parameters which you are supplying, whose value you are giving. So, you are giving these values which are known to you and you want to calculate certain things by which you will be able to design for a better. You know system better queue design and. So, that it can result into good customer satisfaction as well as good output in terms of whatever it be for the company.

So, the input parameters, which we know normally, which we supply is the arrival rate of customers. So, how the customers arrive, how many customers arrive in a certain time. So, that is arrival rate of customers service demand of customer rate at which server works, then arrangement of servers. These things are the input parameters like customer,

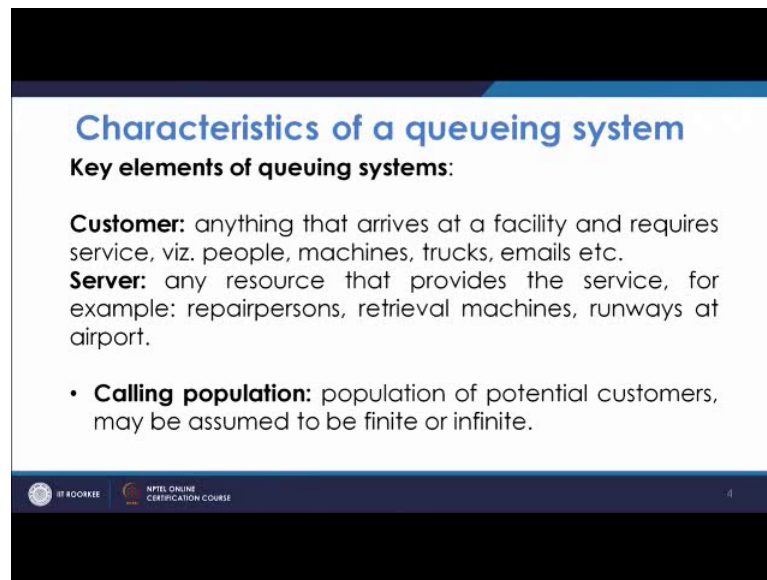
we will serve at what rate, what will be normally the time, in which the server, we will complete the serve service. So, that is service rate. So, similarly, how many servers are there.

So, you have all these parameters in your hand initially. So, these are the input parameters as we discussed that, if you have simple queueing systems, the mathematical analysis can be done and then you can get these output performance measures calculated; however, in normal cases that queueing system is a complex, whenever you have the realistic system, it is not very simple in simple system, you know, you may have a deterministic system, where you just say that the arrival will be at this and this time. So, if we have certain ten customers, you know the time at which they will come and also we know that when I mean for the  $n$ th customer, what will be the service time required, if we know all that then in that case the performance measures can be calculated easily.

So, simple mathematics can do that, but in real case, this is not the situation in real case, the system is complex. We do not know, when the customer is going to come, we do not know what service time will be taken for a particular customer, then there are you know randomness in many aspects. So, for that to take into account simulation is required. So, because simulation is advised, because there are many, you know aspects involved, there will be some deterministic part, there will be some random part. All this is to be taken into account. So, then sees the system becomes complex.

Simulation is the advisable option and it is required characteristic of a queueing system.

(Refer Slide Time: 12:51).



**Characteristics of a queueing system**

**Key elements of queueing systems:**

**Customer:** anything that arrives at a facility and requires service, viz. people, machines, trucks, emails etc.

**Server:** any resource that provides the service, for example: repairpersons, retrieval machines, runways at airport.

- **Calling population:** population of potential customers, may be assumed to be finite or infinite.

III SEM EEE NPTEL ONLINE CERTIFICATION COURSE 4

So, what are the characteristics of a queueing system, what are the key elements as we discussed. The key elements of a queueing system one is the customer. So, what is a customer customer is anything that arrives at a facility and requires service, I mean like people, machines, trucks, emails, all that these are the customers, whichever you know, whatever arrives and it require service emails are coming and they are required to be sent to somewhere by the server, then machines are coming for repair. So, repair person has to repair the machine and then machine will go out, truck will come, you have to load something and then it will go.

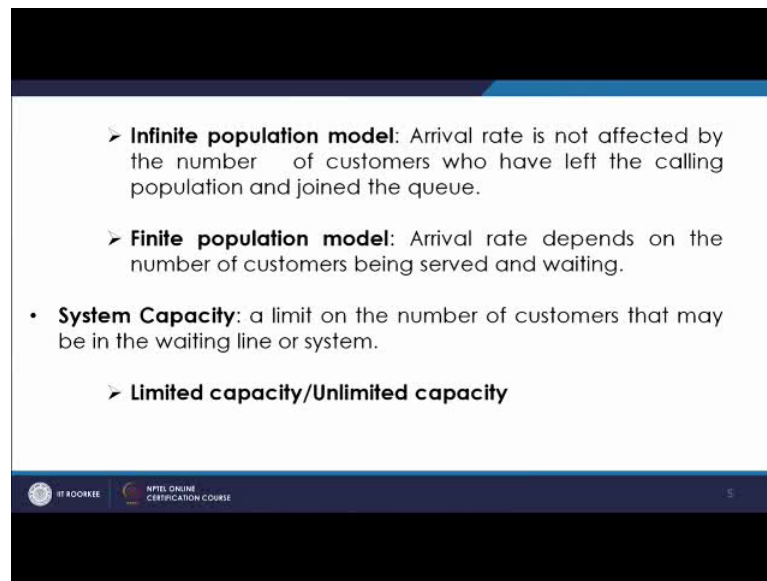
So, these things are customers people used in banks, post office, railway station, and in ticket counters, everywhere. So, you have the persons, these are the customers. Similarly, you have server. Server is any source that provides, the service like the repair person as we talked, if the machine is customer, the repair person is the server, retrieval machines runways at airport. These are the server, if the aeroplane is the customer, which is coming, the runway is the server. Server may be busy server, may be free.

So, server has two status, it may be either busy or it may be you know not busy. It may be idle. So, that is customer and server, then you have calling population. Calling population is population of potential customers means the customer you know outside the system, outside the server, you have a server, you have the customer and the customer has to come from outside. Now, whether it is finite or infinite. Finite means

you have some population, which is finite number of population is there and then infinite means there is no end to it, like in banks is a infinite population customer.

So, customer will be coming. So, there is certain changes, there are certain changes as far as other aspects is to be considered in the case of finite and infinite calling population.

(Refer Slide Time: 15:42)



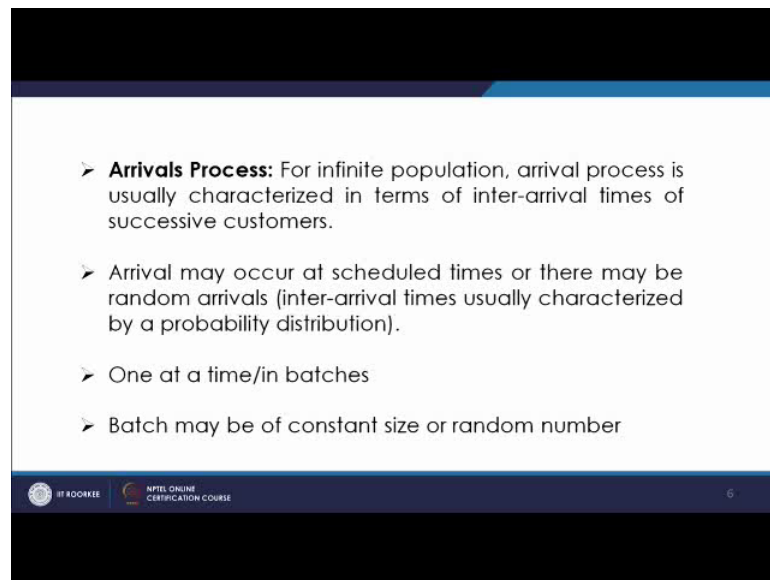
- **Infinite population model:** Arrival rate is not affected by the number of customers who have left the calling population and joined the queue.
- **Finite population model:** Arrival rate depends on the number of customers being served and waiting.
- **System Capacity:** a limit on the number of customers that may be in the waiting line or system.
- **Limited capacity/Unlimited capacity**

IT ROOKIE NPTEL ONLINE CERTIFICATION COURSE 5

Now, what happens in the infinite population model. So, infinite population model arrival rate is not affected by the number of customers, who have left the calling population and joined the queue. Now, in case of infinite population model, the arrival rate will not be affected by the number of customers, who have got the service or who and then also that way it does not depend upon that, where as in the case of finite population, mod model arrival rate will depend on the number of customers being served.

So, as the number of customers, being served, which increasing, that will change this, you know arrival rate. So, that is the difference between infinite population model and finite population model, we will discuss more about it.

(Refer Slide Time: 16:46)



➤ **Arrivals Process:** For infinite population, arrival process is usually characterized in terms of inter-arrival times of successive customers.

➤ Arrival may occur at scheduled times or there may be random arrivals (inter-arrival times usually characterized by a probability distribution).

➤ One at a time/in batches

➤ Batch may be of constant size or random number

IT ROOKIE NPTEL ONLINE CERTIFICATION COURSE 6

So, in infinite, in population arrival process is usually characterized in terms of inter arrival times of successive customers. So, as we see that when we deal with the infinite population model, the customers is coming at one time, second customer will be coming at some time, third customer will be coming at some other time.

So, we talk about the time in between the rivals. So, normally, we talk about it then arrivals also occur in different way. So, that we will discuss. So, before that, we were discussing about the finite population model, in the finite population model arrival rate, will depend on the number of customers being served and waiting like, if you have a large machine, which is to be served by a, you know repair person, once he attends that the arrival rate will be further 0 or if you have a finite population, one person comes certainly, the arrival rate will change whereas, in that case, it does not depend upon how many persons are being served or how many are waiting, because this is infinite population.

System capacity that is a limit on the number of customers that may be in the waiting line or system; so, that is the system come, how many system whether, it can accommodate, how much it can accommodate. So, that may be in the waiting, if you are putting a limit. So, that is limited capacity, if you are not putting a limit, it is a unlimited capacity banks or. So, there are unlimited capacity, anyone can come and join, but in some cases, you can fix how many, can wait how many can come and be in the queue .



So, that is the limited capacity and unlimited capacity like that is for the system capacity. So, we were talking about the arrival process, in case of infinite population. Now arrival may occur either at schedule time or there may be random arrivals. So, we first discussed that arrivals normally, in infinite population models arrival process is characterized in terms of inter arrival times of successive customers.

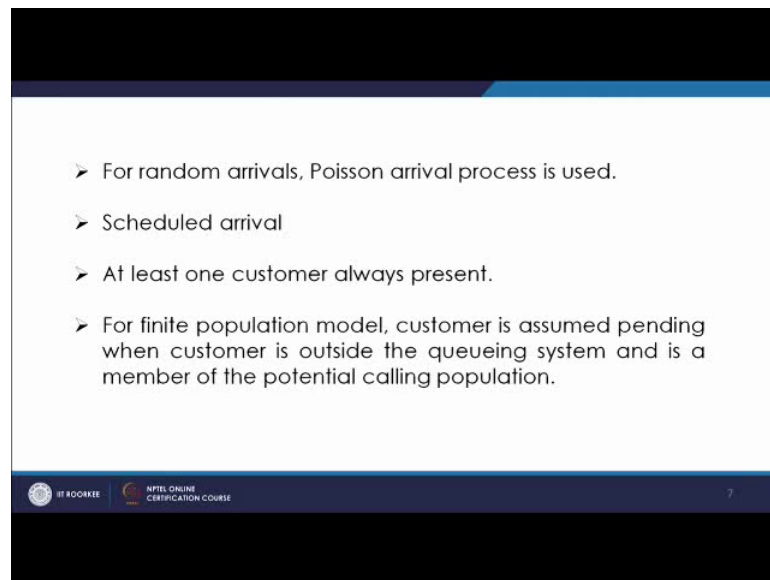
So,  $A_1, A_2, A_3$ .  $A_1$  will be from 0 to 1. First customer, the time taken is  $A_1$  like from first to second. It will be  $A_2$  like that. So, in those cases, normally, you characterize this arrival, you know in terms of inter arrival times of successive customers, then arrivals may be either at schedule time, just like in the cases of doctors or there may be random arrivals. So, as we go to any clinic or the doctor, there we have, we fix the time, we tell that, you have to arrive at that time. So, these are like the example of schedule time arrivals and sometimes, it is very random.

So, the random arrival in that, you have inter arrival times and normally, they are characterized by a probability distribution. So, we have discussed in the last lectures about different types of probability distributions and we have seen that inter arrival times are basically, represented by or characterized by some specific type of probability distribution functions. So, then it may also happen that sometimes the arrival maybe a combination of some deterministic part and some variable part or random variable part. So, as we can say that, at this time plus minus, it may vary by certain time or in that part is not fixed, that is a random process.

So, this way arrival maybe modeled arrival, maybe one at a time or in batches. So, that will be batch arrival single arrival like that, then the batch maybe also, batch maybe of a fixed quantity, a fixed number or it may also be random. So, again randomness maybe, clubbed with the batch strength. So, this kind of arrivals can be designed and basically, these are to be put in the routines. So, when you are simulating the queue at that time, you will have to keep all these things in mind what kind of arrival you want to take.

So, as we have discussed that is the case of random arrivals Poisson process is used, we had discussed that you have the Poisson process, which discusses about the random arrivals, where we talk about the number of arrivals, in certain time interval.

(Refer Slide Time: 22:20)

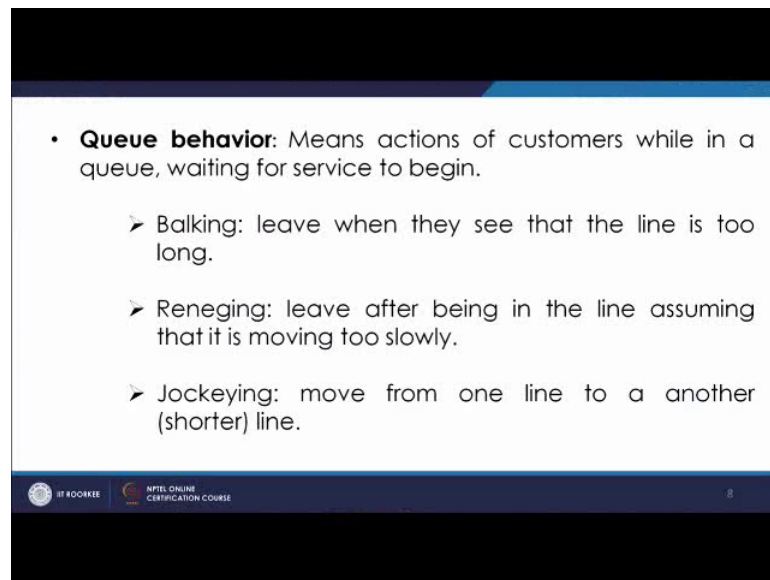


So, as we discussed that the time between the arrivals is exponentially distributed, whereas. So, that is with rate  $1/\lambda$  whereas, in the Poisson process, it is  $\lambda T$  in that time. So, that way it is. So, for random arrivals normally, we go for Poisson distribution, Poisson arrival process.

You may have the scheduled arrival, then there are cases, when we again also assume or we assume that, at least one customer is always present. So, in some cases or the server is never idle like in machine shop, the number of work piece, which is waiting is too many. So, that has to go. So, these are the different kinds of arrivals. Now, in the case of finite population model, the customer will be assumed pending, when the customer is outside the queueing system. So, when he is outside the queueing system, he will be assumed pending, because as the another customer will go, he may get the chance of going into the system.

So, he is a basically, member of potential calling population. So, these are the different, you know aspects related to the arrival of the customers in queue.

(Refer Slide Time: 24:19)



- **Queue behavior:** Means actions of customers while in a queue, waiting for service to begin.
  - Balking: leave when they see that the line is too long.
  - Reneging: leave after being in the line assuming that it is moving too slowly.
  - Jockeying: move from one line to a another (shorter) line.

Now, we will discuss about the queue behavior. So, when a person join is the queue, he has certain actions till I mean or waiting for the services to begin. Now, any customer, who is coming into the queue, he, a feels that he should get a service; so, the typical processes, that the customer will come, he will have a queue or he may not see the queue. He may rightly go to the service, but in normal cases, there will be a queue.

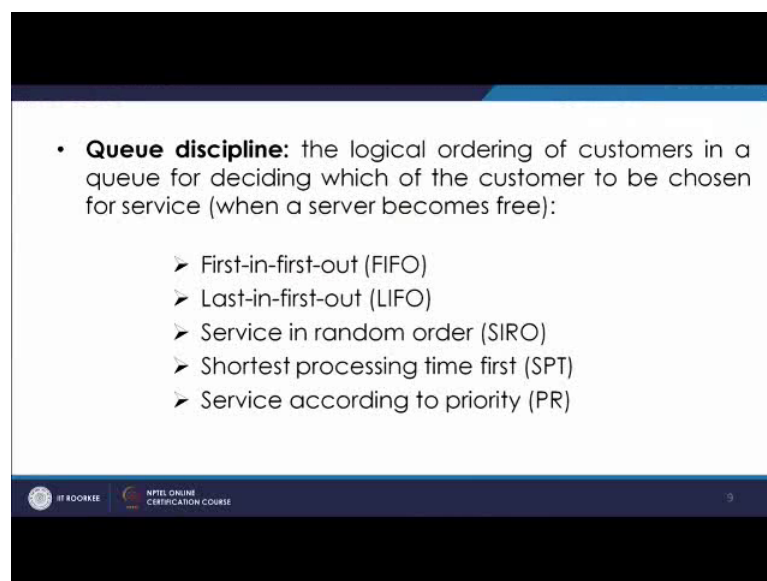
So, he will have to come in the queue and after sometime is spending in the queue, he will go to the counter. So, the time it goes on the counter and starts getting the service, he is waiting time is over. So, he now, he has to wait only for getting the service. Now, the thing is that there are many, you know behavior of the customers because, of which he does different accents, like there is a process known as bulky. So, the balconies that you know tendency of the first customer, because of which he leaves the queue, when he feels that the line is too long. So, maybe it is depends upon the, you know aspect, the way, the person feels, he comes and sees the queue length to be more than certain limit.

In that case, he may go that varies from person to person. So, once he comes before, without joining the queue, he goes in there inters and sees that there is a queue and he leaves the system, he does not going to the queue. So, that is bulking. Similarly, you have reneging. Reneging means the customer, joints the queue and he will be there in the queue for some time and then he assumes that the queue is moving very slowly. So, it will take large amount of time, for him to go to the service counter and get the service.

So, in that case after being sometime in the queue he move sort of the queue and he goes out of the system. So, that is known as reneging, the another behavior of the customer or the action of the customer is related to changing of the line, I mean from 1 to 2. So, many a times the customer is there in one queue and he looks at the other server. So, you have parallel servers and he feels that the adjacent server or any other server, they are the people, are moving at a fast rate. So, he will be going from one cust, one queue to another queue.

So, this behavior is known as joking. Now, as we know you have the queueing parameters based on that, this phenomena may also be modeled. We can said a condition that, if the queue length becomes. So, there may be a case of barking, balking or if the service rate becomes very small in that case, there may be case of reneging and in the case of different servers. If you have different service times by the server; so, at one point of time it may be so that in one of the queue, if it is taking more time, from there it may go to another queue. So, this also can be model. So, there are the queue behaviors.

(Refer Slide Time: 28:32)



• **Queue discipline:** the logical ordering of customers in a queue for deciding which of the customer to be chosen for service (when a server becomes free):

- First-in-first-out (FIFO)
- Last-in-first-out (LIFO)
- Service in random order (SIRO)
- Shortest processing time first (SPT)
- Service according to priority (PR)

NPTEL ONLINE CERTIFICATION COURSE

Queue discipline, the next point is queue discipline. So, what is queue discipline; it is the logical ordering of customers in a queue for deciding, which of the customer to be chosen for service. So, there has to be certain discipline. So, that there is carouse, as we discussed that any person, who is coming for getting service he feels that he should get the service first, in that case there will be carouse, nobody wants to wait . So, how you

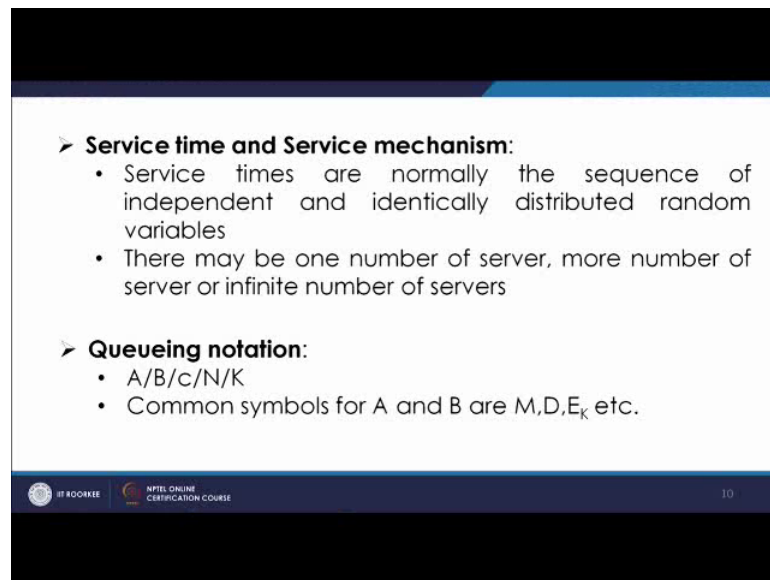
have to decide, who should go first at the service counter. So, for that, there are many cases. You have first in first out that is FIFO. So, also called general discipline.

Normally, whoever is coming first, he will get the opportunity to get the service first. So, we will be the first one to go out, the second person will go second. So, like that, this is the general discipline, that is FIFO, it is known as many a times, you have last in first out, many a times, whatever it is coming and then the service is starts from the end. So, that is known as last in first out, sampling. Whenever, we do, if you are putting the, you know a stocks, in a dark room and then we are taking out may be, then normally, we choose the first one that may be the last sample, which has gone into it.

Servicing a random order that is S I R O. So, in random order, anybody can be picked and he may get the service. So, that is known as service in random order, then you have shorted, shortest processing time first, many a times, we see that which process is you know take, get taking less time. So, normally, we privatized according to that, if we see that some of the process is taking less time, then we try to you know privatize them. So, that is. So, we are going to take that one first, who, which are the shortest processing time and then service according to priority.

So, we are giving priority to certain type of customers. It happens maybe at the doctors or at say many a counters, where you have the Vip's. So, we give service according to prior, there any chances that some Vip's will come. So, we may have to go for P R. So, these are the different kinds of services, you know to discipline, which is normally followed, in the case of queue .

(Refer Slide Time: 31:58)



➤ **Service time and Service mechanism:**

- Service times are normally the sequence of independent and identically distributed random variables
- There may be one number of server, more number of server or infinite number of servers

➤ **Queueing notation:**

- $A/B/c/N/K$
- Common symbols for A and B are  $M, D, E_k$  etc.

IT ROORKEE    NPTEL ONLINE CERTIFICATION COURSE    10

Service time and service mechanism. So, service time as we discussed, it is also the sequence of independent and identically distributed random numbers. So, that is again following certain probability distribution and there may be more number of server or there maybe 1 number of server. There may be infinite number of servers and similarly, you have the queueing notation. So, we will discuss further in the lecture about the queueing notations, the different performance measures. So, in queueing notation, you have normally 6 parameters. We see the 5 1, where A B C. So, it is, this is, one is arrival rate, second is B service rate, C is number of servers and is the system capacity K is the calling population size. So, this way you have the queueing notation occurs. So, in the next lecture, we will discuss about the queueing notations and the system performance measures in general.

Thank you.