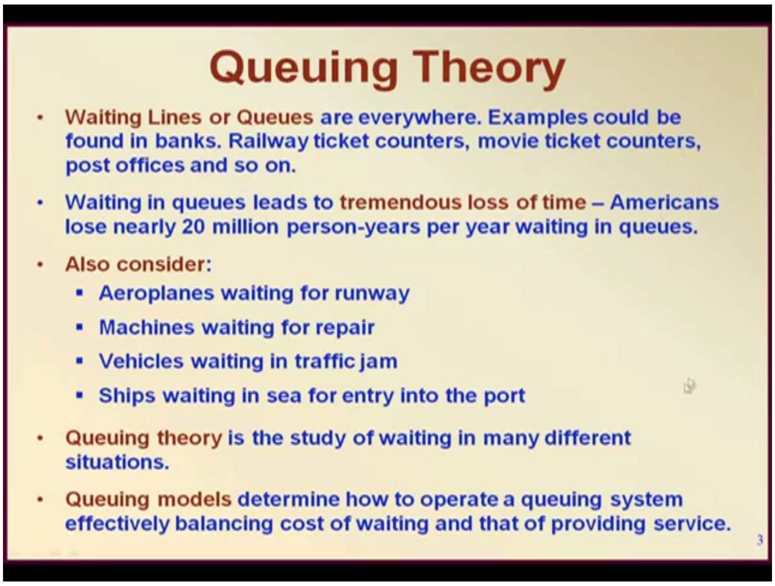
Decision Modelling
Professor Biswajit Mahanty
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur
Lecture-11
Introduction to Waiting Lines

Today we are going to start a new topic that is on Waiting lines which is also known as the Queuing theory. Now waiting lines queuing theory we know the first portion we shall have the introduction. First of all let us know that you know almost everywhere we are having waiting lines or queues. One thing you see particularly in country like ours, most of the situations are such that whether it is a railway counter, the bank counter, the number of arrivals or the people who come for service usually is very large compared to the rate of service, so as a result what happens that a steady state is not reached, right. So if the steady state is not reached then the kind of Queuing theory that we are going to discuss is not very useful. So it must be remembered that one of the first and foremost Queuing theory requirement is that the service rate should be higher than the arrival rate.

If the service rate is higher than arrival rate then should there be any queue, the answer is usually it should not but because there is a probability distribution, both of the arrivals and also of the service, this probability distribution leads to formation of queue at a certain time right at certain time. So that is the essential idea that the foremost forbearing thing about the waiting line of the Queuing theory that we are talking about is that service rate should be higher than the arrival rate. Obviously sometimes we find situations where there are multiple servers that means queue may be single but there are multiple servers, so at that time that all the service rate should be integrated, the total service rate should be found out and that should be higher than the arrival rate so that is the basic requirement.

(Refer Slide Time: 2:45)



Queuing Theory

- Waiting Lines or Queues are everywhere. Examples could be found in banks. Railway ticket counters, movie ticket counters, post offices and so on.
- Waiting in queues leads to tremendous loss of time – Americans lose nearly 20 million person-years per year waiting in queues.
- Also consider:
  - Aeroplanes waiting for runway
  - Machines waiting for repair
  - Vehicles waiting in traffic jam
  - Ships waiting in sea for entry into the port
- Queuing theory is the study of waiting in many different situations.
- Queuing models determine how to operate a queuing system effectively balancing cost of waiting and that of providing service.

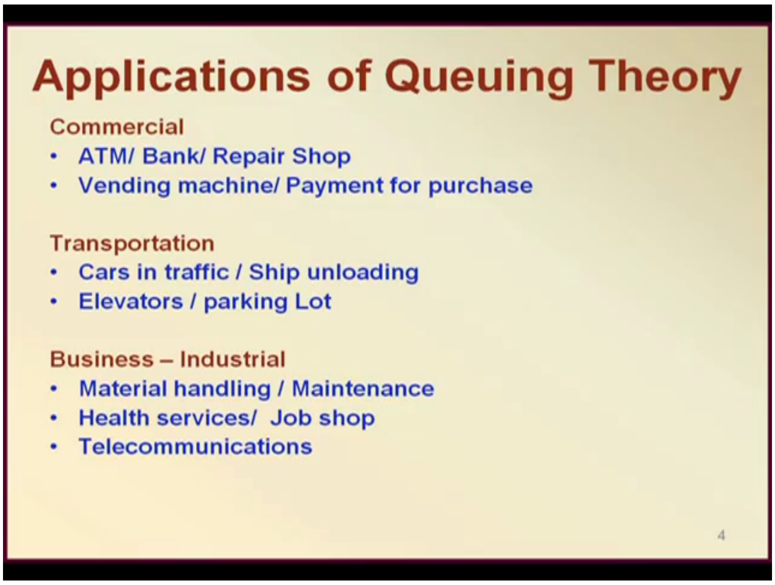
3

This point will also come at an appropriate time, let us begin now. So as you can see here that waiting lines or queues they are everywhere, the example could be found in banks, railway ticket counters, movie ticket counters, post offices and so on. And you know I should be study waiting lines or queues because it leads to tremendous loss of time. One example that I have taken from Hillier and Lieberman's book is, the Americans lose nearly 20 million person-years per year waiting in queues. So all the examples which you saw just now they are all about people, people waiting in banks, ticket counters, post offices and all.

But apart from people there are other kind of things that also wait for example, the aeroplanes waiting for runway, machine waiting for repair, vehicles waiting in traffic jam, ships waiting in sea for entry into the port right. Many of us do not know about the situation the in Port what really happens, very big queue actually forms and sometimes the ships the incoming ones are not even you know they are given you know the go ahead to enter the port until unless the waiting once are already served.

So all these are also leading to huge amount of you know the losses, so the basic idea is about the Queuing theory is the study of such waiting in different situations and the Queuing models determine how to operate a Queuing system effectively balancing on one side the social cost of waiting and the other side is the cost of providing service right. So obviously you know if you do not want the people to wait, you have to spend more money in creating more service facilities, on the other side you know if you spend too much on providing service, the queues will be less but cost will be much higher so a balance or a trade off has to be obtained that is the essential idea.

(Refer Slide Time: 5:04)



Applications of Queuing Theory

Commercial
- ATM/ Bank/ Repair Shop
- Vending machine/ Payment for purchase

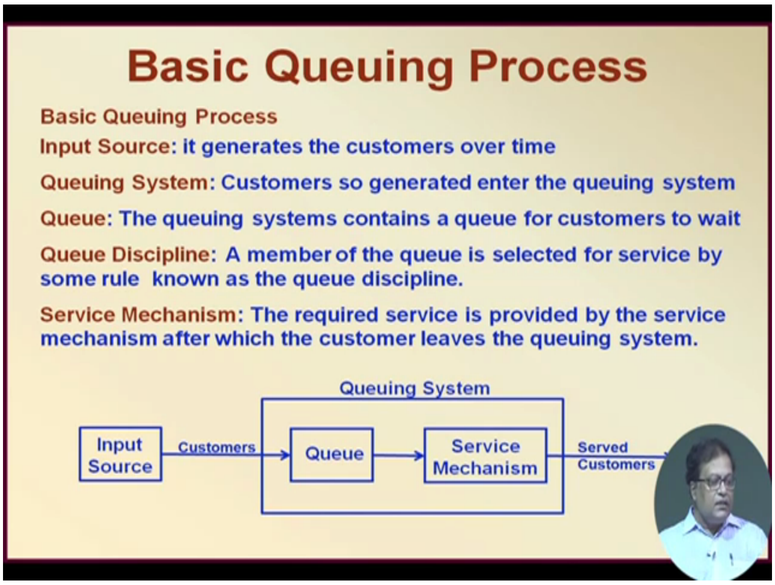
Transportation
- Cars in traffic / Ship unloading
- Elevators / parking Lot

Business – Industrial
- Material handling / Maintenance
- Health services/ Job shop
- Telecommunications

4

Here are some more examples like ATM, bank, repair shop, vending machine, payment for purchase, all these are commercial situations. There are transport situation where there are cars in traffic, there are ships uploading, elevator, parking lots all, in business situations like material handling, health services, jobs shop, telecommunications, so examples are many and however in all the situations the common thread runs that there is a queue that is formed and which we shall show in our next diagram that if you look at this particular diagram that has to be an input source from which the customers come in or you may say that customers are generated.

(Refer Slide Time: 5:51)



Basic Queuing Process

Basic Queuing Process

Input Source: it generates the customers over time

Queuing System: Customers so generated enter the queuing system

Queue: The queuing systems contains a queue for customers to wait

Queue Discipline: A member of the queue is selected for service by some rule known as the queue discipline.

Service Mechanism: The required service is provided by the service mechanism after which the customer leaves the queuing system.

Queuing System

Input Source

Customers

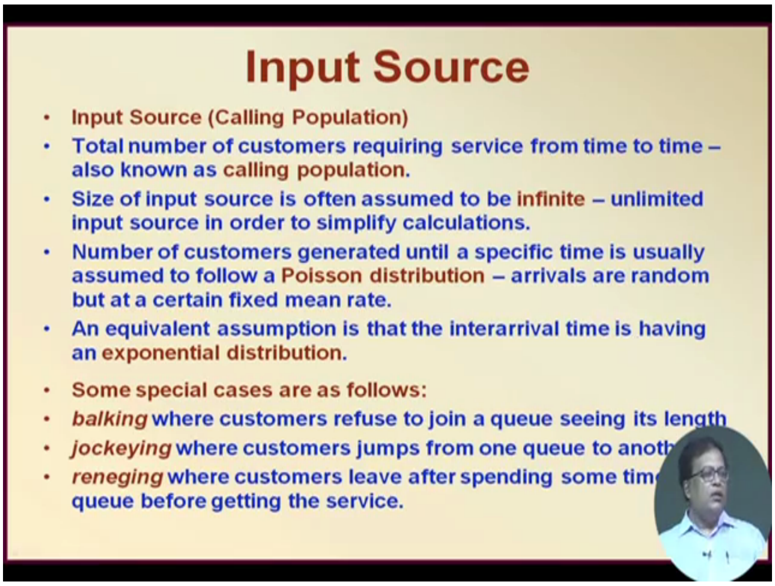
Queue

Service Mechanism

Served Customers

Then there is a Queuing system which basically includes a queue and a service mechanism, the queue where customers are waiting, usually we talk about single queue that is formed and we may also view an infinite length of the Queue, why we will come later. Then that Queuing system should have a service mechanism through which that people the things are served and there is also should be a queue discipline, what is that queue discipline? One example could be FCFS first-come first-served, the first person who comes will be served and finally the served customers they are leaving the Queuing system so that is the basic idea of the Queuing process.

(Refer Slide Time: 6:43)



Input Source

- Input Source (Calling Population)
- Total number of customers requiring service from time to time – also known as calling population.
- Size of input source is often assumed to be infinite – unlimited input source in order to simplify calculations.
- Number of customers generated until a specific time is usually assumed to follow a Poisson distribution – arrivals are random but at a certain fixed mean rate.
- An equivalent assumption is that the interarrival time is having an exponential distribution.
- Some special cases are as follows:
- *balking* where customers refuse to join a queue seeing its length
- *jockeying* where customers jumps from one queue to anoth
- *reneging* where customers leave after spending some tim queue before getting the service.

Now let us look at the details of input source or which is also sometimes called the calling population. See usual assumption about the calling population or the input source is that it is infinite, but suppose it is not infinite, it is a finite source. The moment it is a finite source, then you cannot really expect all the people to come with following an exact probability distribution. For example, you know the most commonly used probability distribution for modelling the arrivals in a Queuing process is the so-called Poisson distribution. The Poisson distribution assumption will not hold if the input source or the calling population is not infinite.

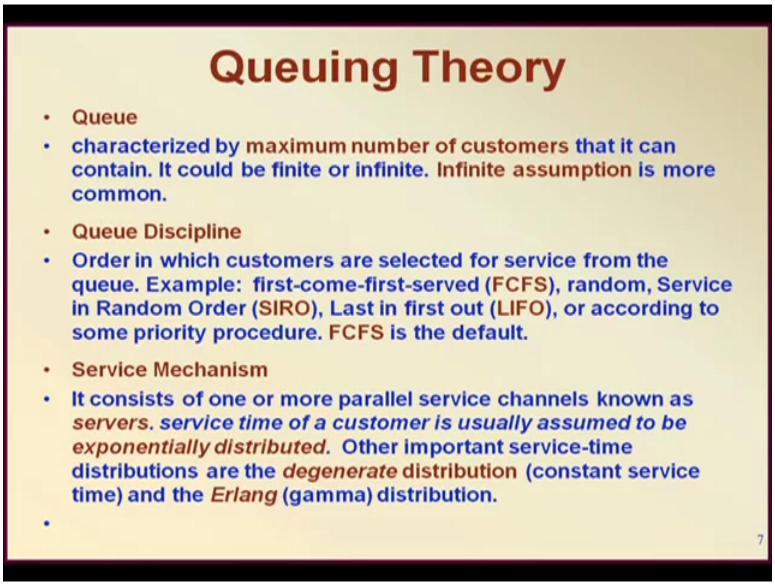
So you can see that if you think that there is a finite population for example, let us look at a class, a class is having some 100 students and all the students after the class is over, they may be going to a coffee shop. So this coffee shop the input population is coming most of the students are probably class, the class has a limited size so for sometimes you can have perhaps say Poisson approximation but not for a longer period because in that case since the input population is limited, the towards the end the assumption of Poisson distribution or a random or a Markovian process will not hold, so these are the issues that must be remembered about the input source.

You see one thing that keep coming back that if the number of arriving customers follow a Poisson distribution which is basically a discrete distribution probability distribution and equivalent is the inter-arrival time which will be a continuous distribution and it is exponential. So there is an equivalence, this equivalent is of great importance but we shall discuss this at an appropriate time. Now while we will talk about input source, there are certain other things which may be called as special cases that also should be remembered. You know the first one can be called balking, what is balking? The customers refuse to join a queue seeing its length.

So it is something like you know there is a highway let us say you know the fuel station where you are moving with your car and you see already quite a number of people are waiting, so you are a customer according to the theory you have been generated right, a customer is generated. But you do not join the queue seeing its length, it is quite long and therefore you go away, but this can be called balking. The other process is jockeying, jockeying is a process where suppose there are multiple queues and multiple servers and one is like normal you see in railway counters, one line is moving pretty fast, the other one is not moving so fast so what you do, you move from one queue to the other right so a new customer is coming who is not generated in the usual process, it is coming from who have already joined another part of the queue that is called jockeying.

There is the third one which is called reneging, when customers leave after spending some time in a queue before getting the service. So you went for getting a railway ticket and you wait in the queue for some time and then you think it is too much, I will come back tomorrow right, so you go away that is reneging so all these different cases can make the Queuing system analysis more complicated and we shall see some of them later on.

(Refer Slide Time: 10:57)



Queuing Theory

- Queue
- characterized by maximum number of customers that it can contain. It could be finite or infinite. Infinite assumption is more common.
- Queue Discipline
- Order in which customers are selected for service from the queue. Example: first-come-first-served (FCFS), random, Service in Random Order (SIRO), Last in first out (LIFO), or according to some priority procedure. FCFS is the default.
- Service Mechanism
- It consists of one or more parallel service channels known as *servers. service time of a customer is usually assumed to be exponentially distributed.* Other important service-time distributions are the *degenerate* distribution (constant service time) and the *Erlang* (gamma) distribution.

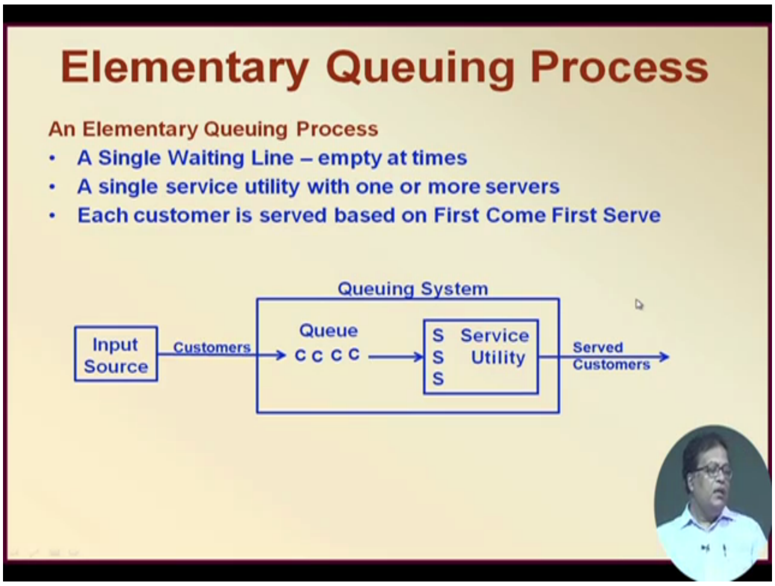
The queue for say as I have told you few minutes back, it is characterised by the maximum number of customers that it can contain, it could be finite or infinite but infinite assumption is more common, why? Because he the moment you think a finite queue space you are actually limiting the analysis and moments you put a limit or a constraint your analysis will be that much more complicated. The next one is the queue discipline, the queue discipline one very interesting thing about queue discipline is that you know the rule or what you call the order in which the customers are selected for service is not actually making the queue you know performance measures any different, so it is something like this.

Suppose you are waiting in front of a doctor, the doctor is following a peculiar queue discipline, what is that let us say last in first-out. So last in first out means what? Means the patient who comes last goes to the doctor first, now it is not something good or usual, you do not like it, as a patient you might have come you know the first in the morning, you are waiting for long and another customer or another patient who comes at the last moment goes before you, you do not like it but what about a doctor? You look at the doctor, to the doctor every patient is the same. If the doctor is seeing let us say the 10 patients per hour that rate remains same. In fact, if the patient or the other people do not tell the doctor, the doctor will not even know that people are not happy because of the queue discipline.

So queue discipline is something which is very important to the customers but not important as far as the queue performance is concerned. So you see all these points come if there are people, suppose there are not people, suppose these are actually jobs, the jobs do not think, since jobs do not think it really does not matter whether it is FCFS or last in first-out or random order or you know something else, so therefore the queue discipline is important at a certain point of time. Obviously even for jobs it may matter, suppose there is something like you know the jobs must maintain a certain temperature, so if you follow last in first-out like in stacking, what will happen there, the job that has come before might have already lost heat by the time it goes in that is not something good.

So queue discipline is an important consideration but unfortunately most of the Queuing performance measures do not change because of queue discipline right, what is the impact of queue discipline we shall again look into later. Then like the customer is generated there is an arrival pattern which is also a service mechanism, the service mechanism consists one or more parallel service channels known as servers right, and the service time is usually assumed to be exponential right or exponentially distributed. Other important service time distribution could be like degenerate distribution which is a constant service time or it could be Erlang or Gamma distribution, so there could be many other distributions as well so these are some common distributions which are important in the context of the queue.

(Refer Slide Time: 14:53)



Elementary Queuing Process

An Elementary Queuing Process

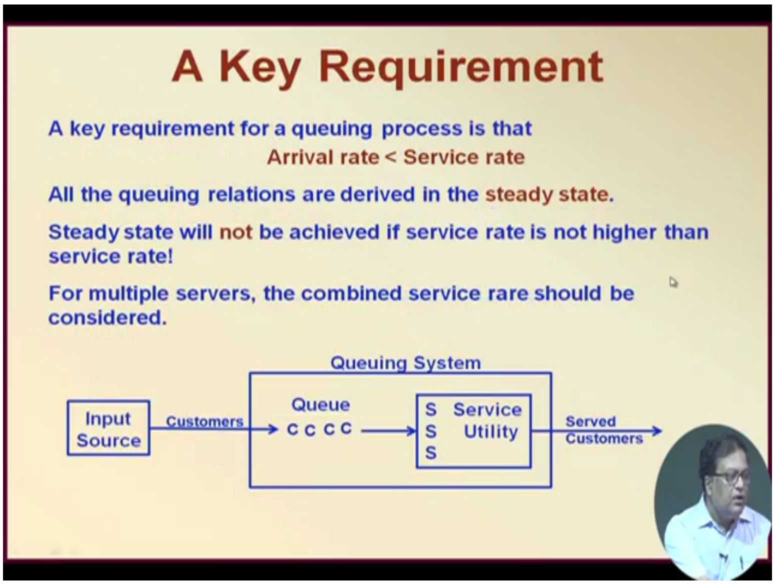
- A Single Waiting Line – empty at times
- A single service utility with one or more servers
- Each customer is served based on First Come First Serve

So here is an elementary Queuing process, if you see that there is an input source, the customers are coming from the input source and they are waiting in the queue and in the service utility there are 3 servers right, so from this common queue the customers are joining one of the servers and thereafter getting the service and after the service is over, the served customers are leaving the system. So once again very quickly, the Queuing process is that there are customers which could be people like people waiting in bank or it could be jobs, it could be aeroplanes waiting for runway or it could be trains waiting for getting into the station, could be a car in a traffic jam, whatever the case maybe it has to be generated.

The customer is generated through a process and that depends on what is the source, input source of the calling population right, the probability distribution has to be there, then it joints a queue most you know typical assumption of the queue is it is an infinite queue length right. Then there is a Queuing system, the Queuing system basically means the queue + the service utility right, so there are some queue characteristics and there are some system characteristics. Usually we do not use the Queuing system word all the time, we simply say system. When we say system, we should basically mean Queuing system right, so there are queue utilities; the queue utilities refer the like L q the number of, average number of customers they are in the queue that is L right.

Whereas, L q the L is on the other side is the total number of customers they are in the Queuing system, so what is the difference between L and L queue, L stands for the number of customers in the Queuing system and L q stands for the number of customers in the queue, so L = L q + number of customers in service, average number of customers in service, so right so these are the things and finally the served customers leave the system. Sometimes there could be situations where the served customers are also waiting in the Queuing system, then the system becomes more complex and analysis will be that much more difficult.

(Refer Slide Time: 17:38)



A Key Requirement

A key requirement for a queuing process is that
Arrival rate < Service rate

All the queuing relations are derived in the steady state.

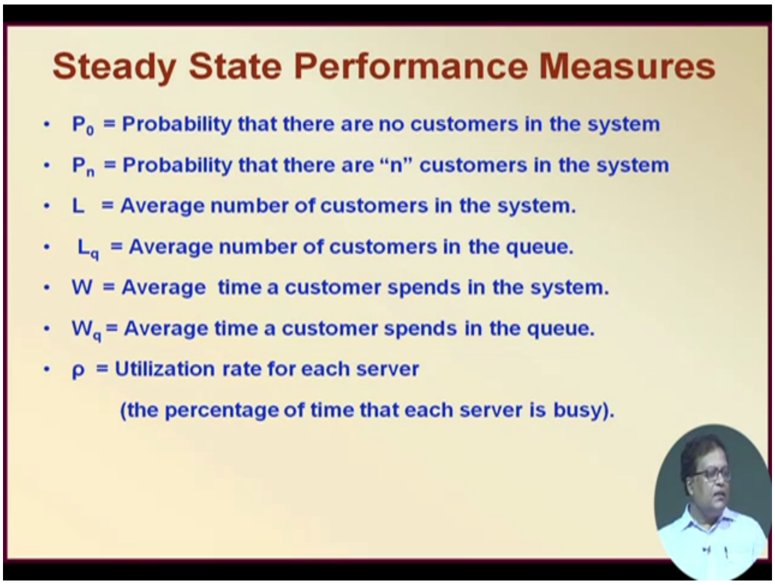
Steady state will not be achieved if service rate is not higher than service rate!

For multiple servers, the combined service rare should be considered.

This point I have already told that there is a key requirement, what is that key requirement? The arrival rate should be less than the service rate right, so all the Queuing relations are derived in the steady state, what is steady state? Steady statements means that the system has come to an equilibrium. Usually what happens that whenever a Queuing system begins, you know there is a transitory period, for some time before it comes to a steady-state assuming that arrival rate is less than service rate, till such time all the queue related parameters are not having stable value, but after some time they reach steady-state and all the queue related parameters they achieve stable values right.

So the kind of Queuing system that we are going to discuss they are all having this particular requirement that service rate is higher than the arrival rate a steady-state is reached and after steady state is reached, all the queue parameters are then considered. As I already said before, for multiple servers the combined service rate should be considered not rare, it should be rate, combined service rate should be considered.

(Refer Slide Time: 19:14)



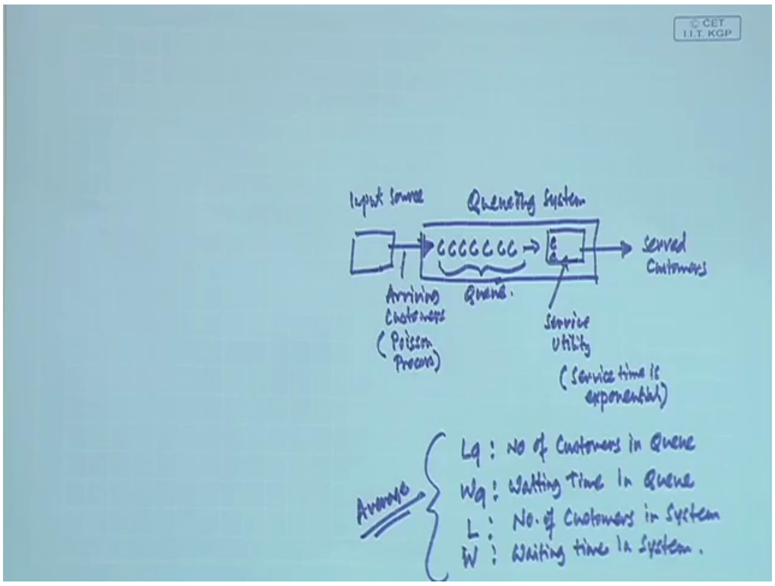
Steady State Performance Measures

- P_0 = Probability that there are no customers in the system
- P_n = Probability that there are "n" customers in the system
- L = Average number of customers in the system.
- L_q = Average number of customers in the queue.
- W = Average time a customer spends in the system.
- W_q = Average time a customer spends in the queue.
- ρ = Utilization rate for each server

  (the percentage of time that each server is busy).

Now I was talking about different performance measures, what are these performance measures let us look at. P 0 the first performance measure, it means probability that there are no customers in the system, P n probability that there are "n" customers in the system, L average number of customers in the system, L q average of customers in the queue, W average time a customer spends in the system and W q average time a customer spends in the queue, finally there is a parameter called Rho which is utilisation rate for each server or in other words, the percentage of time that each server is busy.

(Refer Slide Time: 20:14)



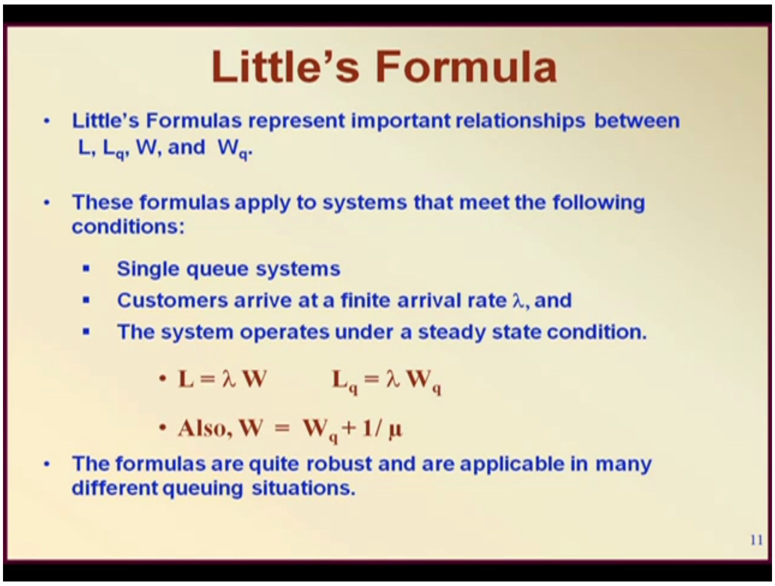
Lq : No of Customers in Queue
Wq : Waiting Time In Queue
L : No. of Customers in System
W : Waiting time In System.

So just let us look at this once again that supposing this is my Queuing system, this is my Queuing system so this is the input source, customers are generated and they are joining the queue here right, and then they go to this service utility and then from here the served customers go away and this one is the arriving customer, so let us look at the usual thing, the **arrival** arriving customers are generated in a Poisson process right, the service utility or the service system or to be more precise the service time is exponential. Not that it cannot be anything else, but this is the most common assumption, so here these arriving customers join the queue and after that they get service, let us say there are 2 servers, so 2 are in service and then they go away.

Now it does not mean the 2 will be in service all the time, suppose there is nobody in the queue there in the system only one person. When there is only one person, there will be only one person in the service nobody in the queue right, so this could be a situation. Suppose there is nobody in the queue and nobody in the service, then the system is idle there is nobody in the queue there is nobody in the service, but other situations there could be nobody in the queue but there could be 2 people in the service that is also possible right, so all these different combinations are possible.

Now this portion is the queue right, so when we talk about measures such as L q and W q right, L q is the number of customers in queue and W q is waiting time in queue right, so this is L q and W q, they refer to the q whereas, L and W they refer to the system, so these are number of customers in system, system means Queuing system and W waiting time obviously all of these are average, waiting time in system right, so these fine differences must be understood please understand that all L q, W q, L, W, they all are calculated after the system has come to a steady-state right. The system has to come to a steady-state then a fixed value of all these performance measures would be obtained then the average value of this because number of customers are changing all the time right all the time what is their average value all right.

(Refer Slide Time: 24:20)



Little's Formula

- Little's Formulas represent important relationships between L, L_q, W, and W_q.
- These formulas apply to systems that meet the following conditions:
  - Single queue systems
  - Customers arrive at a finite arrival rate λ, and
  - The system operates under a steady state condition.
    - $L = \lambda W \qquad L_q = \lambda W_q$
    - Also, $W = W_q + 1/\mu$
- The formulas are quite robust and are applicable in many different queuing situations.

11

So all of this, so what when I say steady-state, I basically mean that average values has come to a steady-state right. So after we look at the steady-state performance measure, is there any relationship that actually relates this variable one to the other right, fortunately there are and this formula is called the Little's formula right there are Little's formulas which actually relate the L and W and L q and W q, so look at these are the relations that L = Lambda bar W and L q = Lambda bar W q right, so these are Little's formula and fortunately for us these formulas are quite robust.

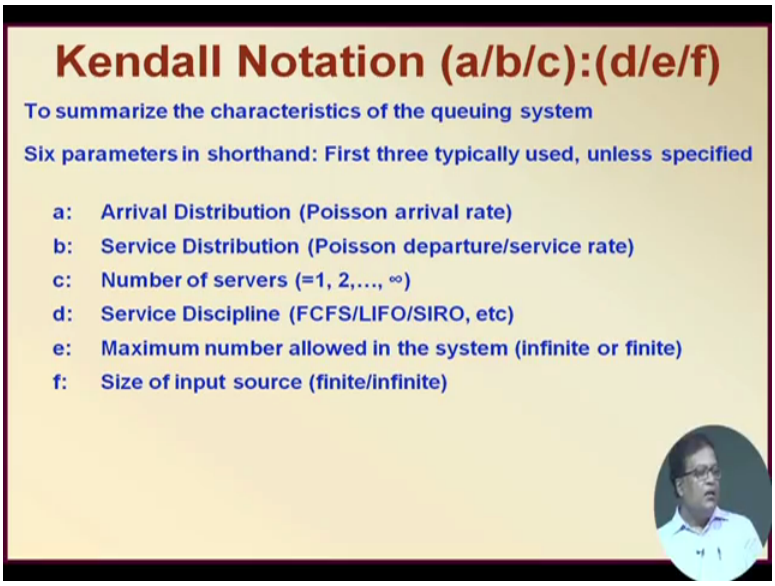
What is meant by quite robust? That means that these formulas are applicable across a large number of Queuing systems large number of Queuing systems, these formulas are applicable. So how does it help us? See lambda bar usually is, what is lambda, obviously I have not said, the lambda is the customers arriving rate right so arriving number of customers per unit time that can be called as lambda. Now you see the lambda could be constant, most of the time suppose we assume it is a perfect Poisson process then lambda is nothing but lambda bar, the number of arrivals are on the you know the average number of customers they are same all the time.

Sometimes but then if the process is not really Poisson for some time right, let us say for initial period nobody comes, after 10 minutes people start arriving then obviously the lambda bar has to be an average lambda of all that is considered that is why it is not lambda really, it is lambda bar. So what is lambda bar? Lambda bar is average value of the lambda or average arrival rate right. In fact, individual lambdas are also averages please remember this, so it is more like average of the averages, average of a particular time, anyhow so that is how they are actually related that is L and L q. Derived formula sometimes also is useful that is W = W q + 1 by Mu. So this we shall see at an appropriate time that if we know only one steady-state parameter out of the 4 that is L, L q, W, W q, using Little's formula we can find out all the other 3.

Suppose we only know L, lambda is known anyway, then we can find W, suppose you know only L q then you can find out W q, from W q you can find out W, from W you can find L right. So these Little's formulas are very important once, I am not really showing their derivation and how they have come but let us take that these formulas are available are very important formulas and these formulas help us to obtain the steady-state parameters of a Queuing system, if we somehow can calculate one of them particularly which one, the formula for which is the simplest.

Suppose the L formula is simplest compute L, W formula is simplest compute W right, so if you can compute at least one of them you can get all the other 3 by using Little's formula, please remember Little's formulas very important. And not only that, they are quite robust that mean they are applicable in many different Queuing situations right.

(Refer Slide Time: 28:36)



Kendall Notation (a/b/c):(d/e/f)

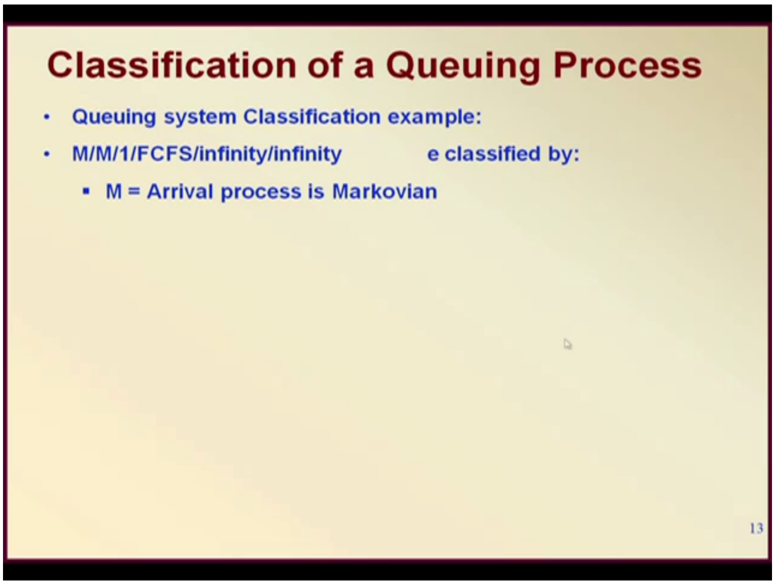
To summarize the characteristics of the queuing system

Six parameters in shorthand: First three typically used, unless specified

a: Arrival Distribution (Poisson arrival rate)
b: Service Distribution (Poisson departure/service rate)
c: Number of servers (=1, 2,…, ∞)
d: Service Discipline (FCFS/LIFO/SIRO, etc)
e: Maximum number allowed in the system (infinite or finite)
f: Size of input source (finite/infinite)

To summarise the characteristics of Queuing system, 6 parameters in shorthand like it is written a, b, c, d, e and f, a is the arrival distribution let us say Poisson, b is the service distribution right that is maybe Poisson departure or service rate in other words exponential service time, c is the number of servers 1, 2, to infinity, d is the service discipline like FCFS, last in first out, service in random order, e is the maximum number allowed in the system right infinite or finite and f is the size of input source right.

(Refer Slide Time: 29:25)



Classification of a Queuing Process

- Queuing system Classification example:
- M/M/1/FCFS/infinity/infinity e classified by:
  - M = Arrival process is Markovian

13

One example of tenders notation, so let us look at MM1, FCFS, infinity, infinity right. Now that this classification is M is the arrival process is Markovian, Markovian means either Poisson or exponential so arrival process is Poisson. Number of arrival is Poisson, enter arrival time is exponential, service time is exponential or you know the number of service which is normally not used that is what is known as Poisson. Number of servers 1 because MM1, then FCFS is the queue discipline that is the first-come first serve, the system size is infinity in infinite waiting line that means the queue length is infinite and then last infinity is population size that means the input population size or the calling population size.

Then notation, M stands for Poisson arrival or exponential service time right. Now there could be others like D deterministic constant arrival rate or service time or G for general right general probability for arrival or service time. So this is about introduction, we stop here and from our next class we see in details, thank you.