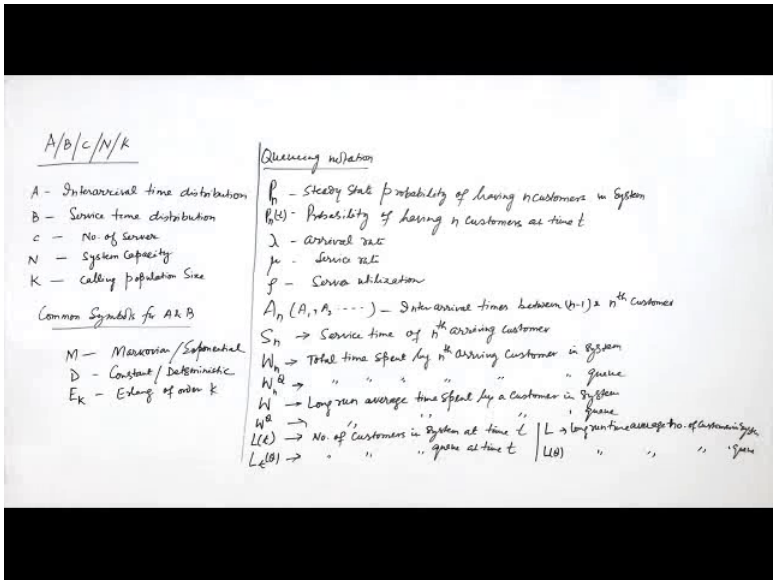
Modeling & Simulation of Discrete Event Systems
Dr. Pradeep K Jha
Department of Mechanical and Industrial Engineering
Indian Institute of Technology, Roorkee

Lecture – 12
Performance Measures of Queueing System

Welcome to the lecture on performance measures of a queueing system. So, in this lecture, we will talk about the different performance, measures in a queueing system before that, in the last lecture, we will discussing about the queueing notations and as we discussed that, there are normally 5 or 6 terminologies, 5 or 6 terms. Normally, we will discuss about 5 terms, which are used in the queues and these terms are known as ABC NK.

(Refer Slide Time: 00:53).



So, this is the normally notations, which are used in queues and as we know that this A's normally, we stock talks about the arrival rate distribution or inter arrival time distribution. So, A will be inter arrival time distribution, then B will be the service time distribution and C is the number of servers and N is basically, the system capacity and K is calling size, calling population size. So, as we discussed, that you have these five terminologies, this will be coming normally.

So, that is A B C N K inter arrival time distribution is normally it is Poisson distributed interval time is basically, then exponential distributed, than arrival number of arrivals

will be Poisson distributed. Similarly, you have Poisson process, similarly, you have B service time distribution, that is how much. So, this is basically, talks this basically, talks about the time between N minus 1 and N customers, if you, we talk about A N A N means the time between N minus first and N th customer.

So, like that similarly, B N B N means, it is normally, the service time for the N th customer, which has arrived C will be number of servers. So, you may have 1 number of server, you may have more number of servers, multi server, single server system capacity. It may be finite or infinite and the calling size population, column population size will be finite or infinite. So, normally, what is done is the when this is following that Poisson distribution or exponential distribution, I mean depending upon what we take in. If it is the inter arrival time, exponential distribution, if the number of arrivals in certain time, then it is again the Poisson process.

So, that is basically said to be a Markovian type of, you know distribution and it is represented by M. So, M, this will also be known as M and this is if the server is 1 in that case. It is MM 1. So, MM 1 is normally the Q, which talks about, you know Poisson distributed service, through Poisson distributed arrival and service and then C is the server that is 1 . So, that it is normally, the queue notation and this two will be normally infinity. So, that infinity, no need, not be returned every time.

So, this way a queue is noted denoted now. So, common symbols for A and B, as we discussed are M is Markovian or exponential D. We also talk about D, when it is deterministic or fixed. So, it will be constant or deterministic. Similarly, you may have Erlang distribution with K th order. When it happens in phases, you may have other also, you know kind of symbols for A and B, but normally, we go for Markovian in most of the cases. So, when it is Markovian arrival and service rate and there is one server, it is M M 1 and if both are infinity N N K, then M M 1 infinity and normally, it is represented by M M 1.

Now, you have certain queueing notations, which we must know, you have some terminologies, which we will come across, when we talk about the queues. So, among them 1 is P N. So, when we talk about P N means, you want to see the probability of having N customers, in a steady state system. So, it is steady state probability of having

N customers in system. So, this is basically, the steady state probability values, then if we talk about P N T, it means this talks about a time.

So, it is the probability of having N customers at time T. So, this is the one next is you have lambda. Lambda is the arrival rate, similarly, you have NU that is service rate. So, this talks about, also you know service rate, then comes RHO, that is server utilization, you know we are also interested to know as we discussed that for what fraction of time the server was busy or it was idle. So, that is basically, you know calculated using this factor value server utilization.

Then you have. So, as we discussed A N means A 1 A 2. These talks about the inter arrival time between N minus 1 and N th customer. So, as we discussed that the arrival may be given in terms of the number of arrivals in certain time or it may be given as the inter arrival time between the customers. So, if the inter arrival time between the customers are given in that case, the time of arrival can be computed by adding those times whatever I have happened earlier.

So, that way when arrival time can be calculated; so, A N that is A 1 A 2, these are the inter arrival times. Similarly, S N it is used for the service time of N th arriving customer. So, S 1 will be the time required for the first serve, I mean customer to get the service that is S 1. Similarly, S 2 is for the second customer, to get the service that is time required. So, that this is given by that may be a fixed one quantity or that may be given by a probability distribution function or. So, next is you have W N. So, this is known as total time is spent in the system by N th arriving customer in system. So, it will be W N is total time a spent by N th arriving customer in the system, in the system means as long he is getting, the even the service.

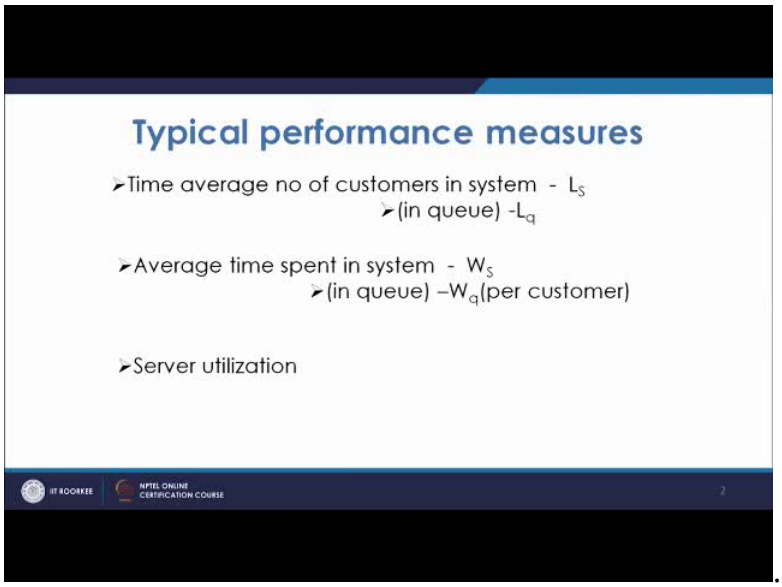
So, that is the total time spent by that N th arriving customer and if we I take W N Q, when we put this Q that basically, is the same thing, but it will be in the queue. So, it will be total time spent by N th customer in queue. So, it does not take the time into account, when the customers joins the queue. So, it will talk about the time till it is before the customer, who is getting the service, once that customer departs then he joins the queue. So, this is basically, the waiting time for the customers. So, that is you can also till like the waiting time for the customer N th customer in queue and waiting time, for the customer in the system that is S.

Similarly, you have another 1. So, when we talk for the N th customer that is W N and when we take W that is long run average. So, long run time average, when we talk about. So, as and it is standing towards infinity. So, that is known as long run average time a spent by a customer in system. So, when we talk about the long run average time is spent by a customer in system that is the W.

Similarly, you have W Q. So, that will be for same thing, but when it is in queue, the next, parameter which is that is L T L T is the number of customers, in the system at time T. So, and then. So, this is a number of customers, in system at time T and way if you talk L T Q. So, that is basically, number of customers, in queue at time T. So, if you talk about L and L Q L will be average number of customers in system. So, that is long run time average number of customers in system and long run time average number of customers in queue that will be L Q. So, this way you have L and L Q.

So, you have when L or you may have L Q that is long run time average number of customers in system and similarly, you have long run time, average number of customers in queue. So, this way threes are you know the notations, these are terminologies, which we will come across, when we discuss about the queueing behavior. So, we will be finding certain values, we will be having certain values and you will be finding certain values, when we analyze, the queueing system. Now, we will move to the typical performance measures in a queueing system.

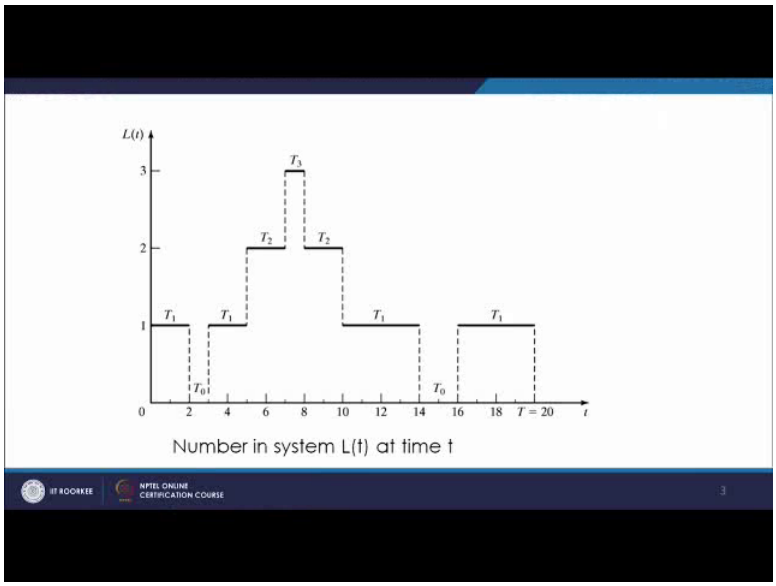
(Refer Slide Time: 15:07)



So, the typical performance measures, which are normally, we are interested in will be time average number of customers in the system, if it is time average number of customers in the system. It is a less and if it is in the queue. It will be L Q, similarly, average time spent in system, that is W S and average time spent in the queue will be W Q and the server utilization. So, how they are basically, you know calculated. So, that can be seen by looking at. So, as we know the time average number of customers in the system or average time spent in the system; so, any person who is spending, how much of time and total time if you know you divided. So, that will be that way it is calculated.

So, we will solve for a problem, which is then it will be more clear that how it is computed.

(Refer Slide Time: 16:22)



So, let us see this is the number in system L T at time T. If you look at this graph, here what we see is it, is in the system. So, at time T; so, what we see is from 0 to 2 seconds or 2 minutes, whatever it be the unit B, you have 1 number of person in the system. So, he came at 0 time and in the system, there is one person. It is means that there is no person in the queue.

So, he has come and he has joined the queue. So, it has not to wait in the system he has gone and then after that in the system, there is no one. So, it has going to means this person, who has come, he has left the system at time 2 seconds or 2 minutes. So, let us say, it is taking minutes. So, in 2 minutes, he is going out of the system, then again in the

system. You have the number one at time 3 minutes, then at time 5 minute, this number goes to 2. It means the person has come here. So, second person has come here, second persons arrival is at time 3 minutes.

At this point at 5 minutes, the third person has come and at this time, you have at 7 minutes. You have the third person coming, second, third and fourth person is coming. So, at this point you have number of persons, in the system is 3. So, at this point, if you look at from this time to this time from, for this duration, you have 3 persons, in the system at this point, one person is coming out of the service. So, this is the second customer, because this customer, who has come here and if we assume that this is a general discipline, first in first out in that case, the second customer will leave this system at this time.

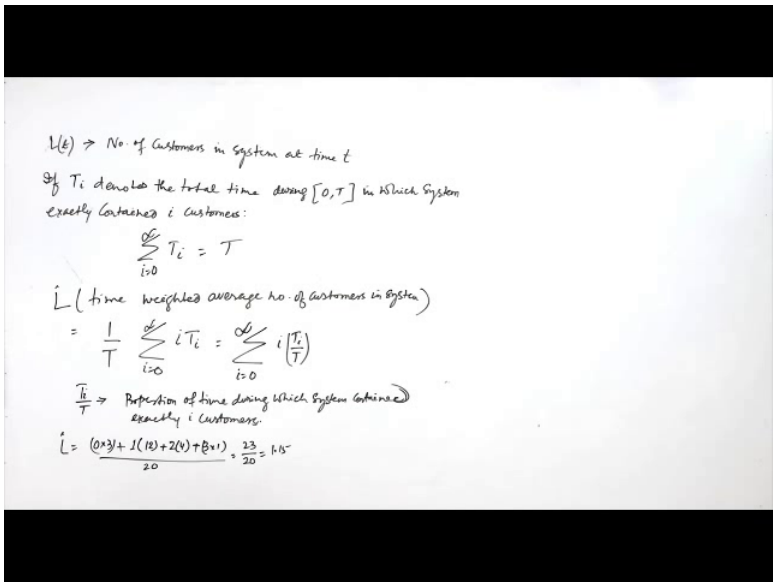
Then the third customer is going to leave at this time, forth customer is going to leave at this time fifth customer is coming at this time and then sixth customer is. So, fifth customer sixth customer, fifth customer is going at T equal to 20. So, this time he is going. So, what we see is that the server is busy, since beginning up to 2 minutes from 2 to 3 minutes. It was not busy, then we were it was busy from 3 to 14 minutes from 14 to 16 minutes. It was not busy then from 16 to 20, it was again busy.

So, as long as the server utilization is there out of 20 minutes. We see that for 3 minutes, it was not busy. So, utilization of server is 17 by 20, that is 85 percent. Now, the, thing is that this graph tells us about the number of customers in the system, but there is another parameter, which is to be calculated that is number of customers in the queue. So, from this graph, we can find how many numbers of customers are there in the queue. So, number of customers normally, in the queue will be L T minus 1.

So, normally, even it is 1 the number of customers, in the queue will be 0. So, the number of customers in Q L Q that will be 0 from 1 to 2 0 to 2 minutes. Similarly, it will be 0 from 3 to 5, but it will be 1 in this period. It will be 2, in this period. It will be N 1 in this period. It will NE again 0 in this period, it is anyway 0 in this period and it is again 0 in this period. So, that is the number of customers in system not in system. It is in queue. So, this is in the system and when we try to find the L Q that time, it will be lesson by 1. So, this way this quantity, is to be calculated.

Let us calculate, what will be the average number of customers in the system. So, we will find time, average number in system. Now, let us first find the expression for these quantities.

(Refer Slide Time: 22:15)



So, if L T L T, that basically denotes the number of customers in system at time T. So, now, we have the time horizon up to 20 minutes. Now, if T I denotes the total time during 0 T in which system exactly content I customers.

So, this T I that is the time duration for which; so, 0 to T, now, if T I is that time duration, during which the system had I number of customers. So, in that case I into T I upon T. So, that will be now. So, what happens first of all summation of T I I from 0 to infinity that will be T; so, in that case l, this will be that is time weighted average number of customers, in system, it will be nothing, but 1 by T summation of I equal to 0 to infinity I T I. So, time weighted average number of customers, in the system, it will be I T I. So, if for T I duration, you have I number of customers and then we are basically, averaging because of the total time from the total time.

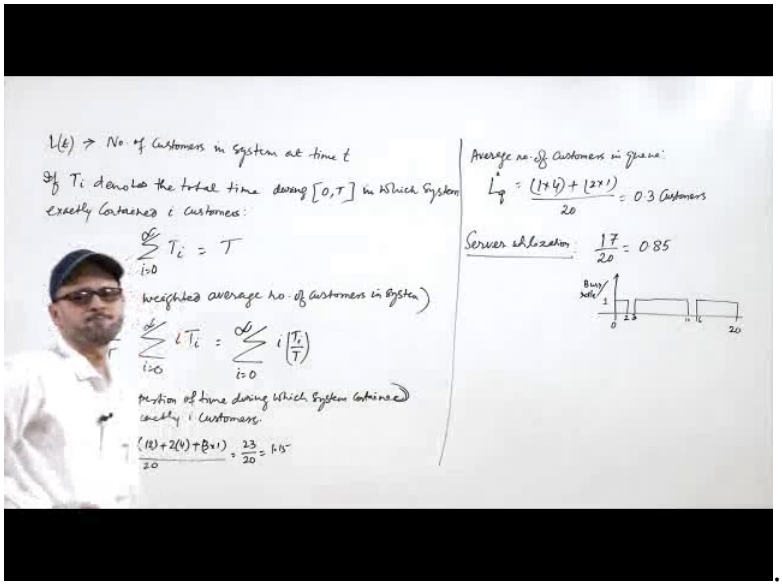
So, it is 1 by T. So, this way you can get the average number of customers, in the system. Now, you can further find this because T being constant, you can write it as I T I upon T. So, this way T I upon T is the proportion of the time, when system contained exactly I customers. So, this T I by T, what is this time, this time is that proportion of time, during which we have number of customers equal to I.

So, this is proportion of time during which system, content exactly I customers. Now, if we try to find using this formula average number of customers in system, then we can find from this graph. So, what we see is you have 0 number of customers, between 2 to 3 and 14 to 16. So, I when it is 0 T I will be 3. So, it will be estimated value we are finding. So, that will be 0 into 3 plus, you have one number of customers in the system, for how much time this is 2 plus, again 2 4 plus 4 8 plus 4 12.

So, you have 12 custom. 12 you know time units and at that time, you have one number of customer in the system. So, I is 1 and this is 12, then we go to I equal to 2 I equal to 2 will be from here, it is 2 plus 2 4. So, two number of persons into 4. So, it will be 2 times 4 and similarly, you have 3 customers, I number of I is 3 for 1 minutes. So, 3 into 1 and whole divided by T. So, T is 20. So, this is coming out to be 12 plus 8 20 plus 323 upon 20 that is 1.15.

So, 1.15 the average number of customers are there in the system, if we talk in the queue, if we talk average number of customers in queue. So, what will happen in average number of customers in queue.

(Refer Slide Time: 28:08)



When we talk about, average number of customers in queue, as we discussed average number of customers, in queue, in that case that value L Q will be basically, 1 less than this value. So, this will be again 0 it will be 1. So, 0 into. So, for this, it will be 0, this will be 1, this will be 2.

So, 1 into 4 plus 2 into 1 . So, it will be L Q and that will be 1 into 4 plus 2 into 1 by 20. So, 0.3 customers. So, this is how the average number of customers, in the queue can be found, then we can also find, the you know server utilization as we discussed, that by looking at this graph or this figure, you can find, the server utilization and server utilization, will be basically, you know 17 minutes divided by 20 minutes. So, it will be 0.85.

Now, we can also find the average time is spent by the customer, in the system. So, average time spent in the system by a customer. So, per customer, we can find the average time spent by a customer. So, the second parameter that is server utilization. So, server by utilized it was not utilized from 2 to 3 and 14 to 16. So, it was utilized for 17 minutes out of 20 minutes. So, it is 0.85. So, that basically, you can find it from a graph and on that in the, on the, this time horizon you have something like this, if we take it as if the one indicates.

So, you have 2 to 3 is. So, 2 to 3 was it was idle. Similarly, you have from 14 to 16, it was idle. So, it was 14 to 16, it was idle. So, this way this fraction from here you can find that the, what is the time the server is busy or it is idle, then another parameter which can be found maybe that you know time is spent by the customer average time spent by the customer. Either in the system or in the queue. So, that is basically, you can in the system, how much is the time spent that you can see that how long he was there in the system and then it can, you can divided further by the number of customer.

So, per customer, you can find similarly, in the queue, how much he has a spent, how much time, he was in the queue, you have to remove the time for which he was there in the, on the server. So, this way you can find those things, maybe we can discuss about this, you know findings, in the next course of lectures. Where we will discuss about, this performance measures calculation and also we will see, how they are programmed, how they are stimulated, using the computers.

Thank you very much.