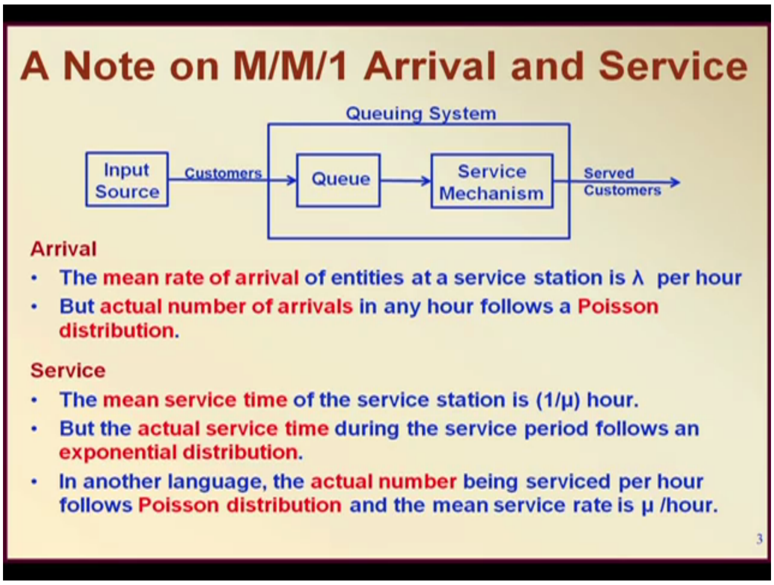
Course on Decision Modeling
Professor Biswajit Mahanty
Department of Industrial and Systems Engineering
Indian Institute of Technology Kharagpur
Module 03
Lecture No. 15
Queuing Examples

Today we begin with setup queuing problems, in our previous classes we have seen what is the importance of queuing theory and apart from that we have also checked on the basic birth and death process and there after we have seen that basic derivations for MM 1 queues. And now it is also we did some problems of MM 1 but we to do some more problems to really clear our concepts on the queuing theory, particularly on the most important queue that is that single server Markovian queuing process.

(Refer Slide Time: 01:02)



Now basically just at this simple diagram we have already shown it before that there is an input source from where the customers arrive join a queuing system, the queuing system consists of 2 parts the first is the queue and the 2nd is a service mechanism. And from the service mechanism where there are 1 or more servers is the served customers leave the queuing system.

1 of the most important thing to understand particularly for the very basics MM 1 process is that the meaning of arrival and service distributions. You see when we tell the arrival is 6 per minute you see it is not really possible that the arrival is a constant 6 per minute. This 6 per

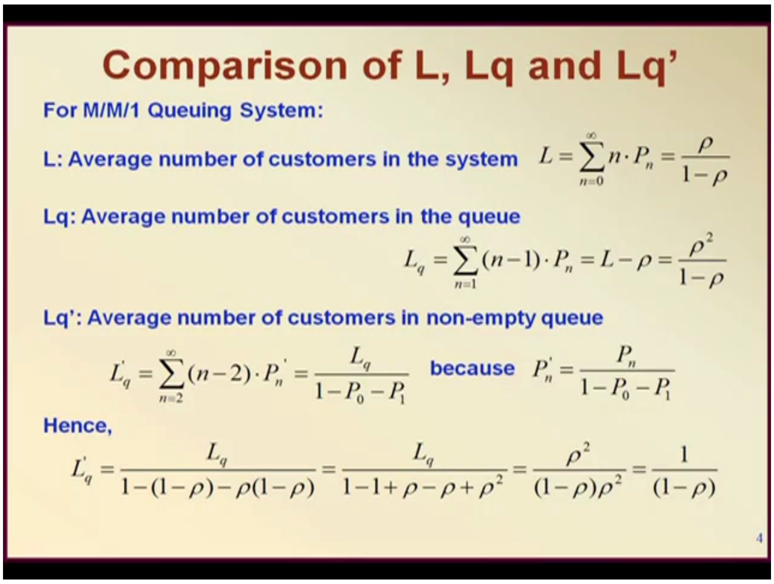
minute is essentially is a mean rate of arrival, is it alright? Mean rate of arrival of entities to the service station is let us say Lambda per hour.

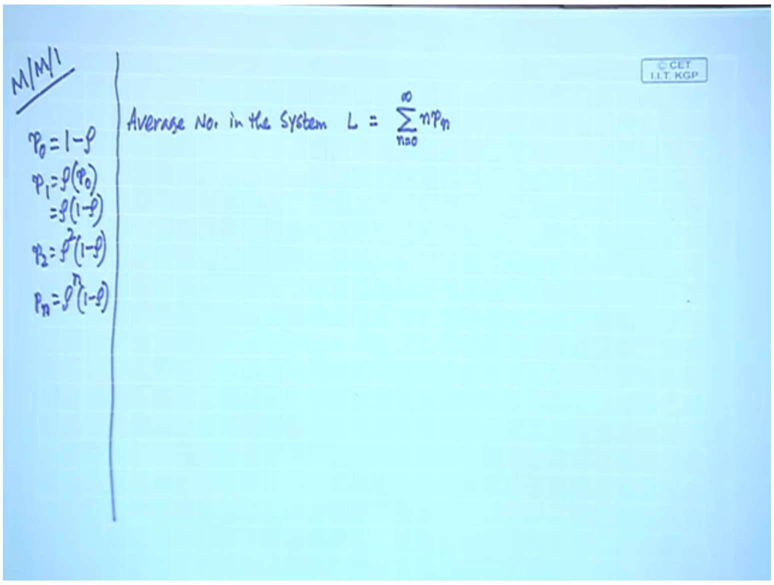
So what is the distribution? The actual number of arrivals which will not be exactly 6 per hour or whatever is the value of the Lambda at that point it is actually following a poison distribution, right? So this point has to be very clear in the problems sometimes we may be writing the arrival rate is 6 per hour following a positive distribution, what exactly it means? That the actual arrivals is 6 per hour around that value is it not? It may not be exactly same but the mean value is 6 and this actual arrivals following what is known as a poison distribution.

Service time on the other hand usually is expressed in the form of exp1ntial distributions. So, what exactly it means? Is that the actual service time during the service period follows an exp1ntial distribution, what is the mean of it? The mean service time of the service station is 1 by Mu per hour. Earlier also I have said and we shall see it further today again that if the actual number being service per hour you know if the service time is following exp1ntial distribution the actual number being service per hour follows poisons distribution with a in service rate of Mu per hour, right?

So once side is the service time, which is exp1ntial, which is equivalent to the number being serviced is having a poison distribution the same logic can be extended to arrival also that if the number of arrivals in a given time follows poisons distribution then enter arrival time follows exp1ntial distribution, right? So these are the basic Markovian process and we must remember that, so it is a kind of review.

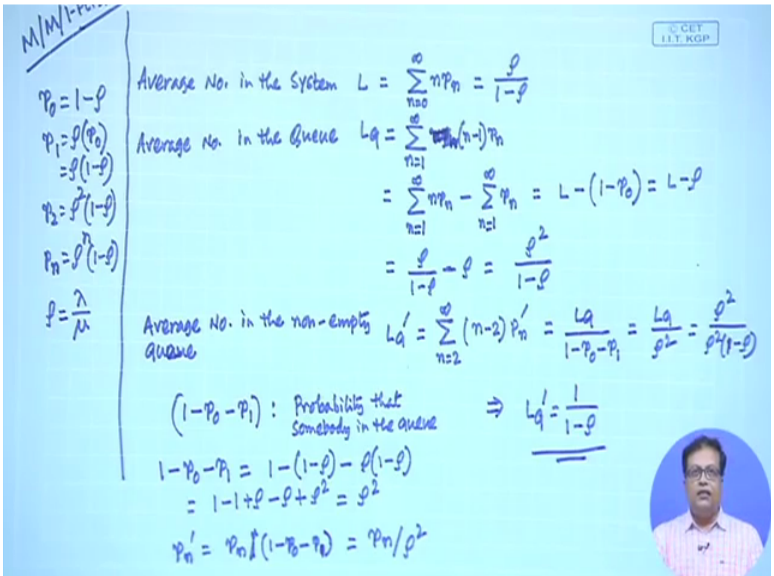
(Refer Slide Time: 04:20)





And now let us see 3 very important parameters the L, the LQ and your LQ dash, right? Three very basic parameters of average number of customers with regard to the MM 1 system. The first 1 as you can see is we have already derived that is average number in the system sometimes it can also called the expected number is L which is defined as some over n equal to 0 to infinity because upto infinity is you know that is the queue space that we can have multiplied by n Pn, right, where Pn is the probability values.

(Refer Slide Time: 05:35)



Please recall for an MM 1 queue we have got some basic results, what are those busy results? Already we know the know that P 0 equal to 1 by rho, P 1 equal to row times P 0, equal to row times 1 minus rho, P 2 equal to rho square 1 minus rho, generalizing Pn equal to row to the power n 1 minus rho, right. So all this basic results are already known with us and we also know that rho that traffic intensity is Lambda by Mu, where Lambda is rate of arrival and Mu is the rate of service.

So this basic results for MM 1 FCFS infinity-infinity queuing system this basic results are already known. So if this basic results are already known, the first parameters that is average number in the system L equal to some of n equal to 0 to infinity it is an expected value of the number, right. Since it is an weighted average that is why the individual ends are multiplied by specific P values. What is the value of this? The value this is rho by 1 minus rho, right, that is the average number in the system L.

The 2nd parameter that is average number in the queue, right, average number in the queue is LQ, what could be LQ? LQ you see at any point of time 1 person maximum could be in the service, because there is a single server, is it alright? So supposing there are only 1 person in the system that person will be in the service, if there is nobody then obviously the question does not arrive arise. But supposing there are more than 1 person then that person will be in the service, if there are 2 person then 1 person will be in the service and 1 person will be in the queue.

So you can understand that LQ now will be having a sum of n equal to 1 to infinity then n Pn, right. Not really n Pn basically it should be n minus 1 into Pn, why? Because n is the number in the system and n minus 1 is the number in the queue, is it alright? So since n minus 1 is a number in the queue this term can be simplified be written as n minus 1 into infinity n Pn minus sum over n equal to 1 into infinity Pn, right. Why n minus 1 once again? Because n is the number in the system, n minus 1 is number in the queue.

Now look here n equal to 0 times P 0 that is 0 anyway. So since n equal to 1 and n equal to 0 sums both becomes exactly same. And what is this sum n equal to 1 to infinity Pn this is nothing but 1 minus P 0, because this is all the terms other than P 0. So this is nothing but L the first term is nothing but L and 2nd term is nothing but rho, because it is L 1 minus P 0 and look P 0 equal to 1 minus rho, so 1 minus P 0 is equal to nothing but rho, so this term becomes L minus rho. So what is L minus rho? It is rho by 1 minus rho 1 minus rho equal to rho square by 1 minus rho, so this is the expression for LQ rho square by 1 minus rho.

Now there is a 3rd term which we also like to discuss in this context which is the average number of in the non-empty queue. Sometimes this also becomes important parameters average number in the non-empty queue which we called as LQ dash . Now what is non-empty queue? Non-empty queue is let us say that in a particular queuing system there say 10 people you know in the queue at a given point of time, but all of them get serviced and some more people joined, some people again leave and everything at some part of time we will find that there is nobody in the system, right.

So let us say out of 1 hour that were somebody in the queue for some time, there is some other time when there was nobody in the queue, what would be that value? That value will be 1 minus P 0 minus P 1, right. Why? Because look here what is the significance of 1 minus P 0 by minus P 1. You see this is the probability that somebody in the queue, why? Because you see P 0 means system is empty nobody in the system, P 1 1 person in the system and that 1 person in the system is you know really is in the service, because there is a single server and the person must be in the service.

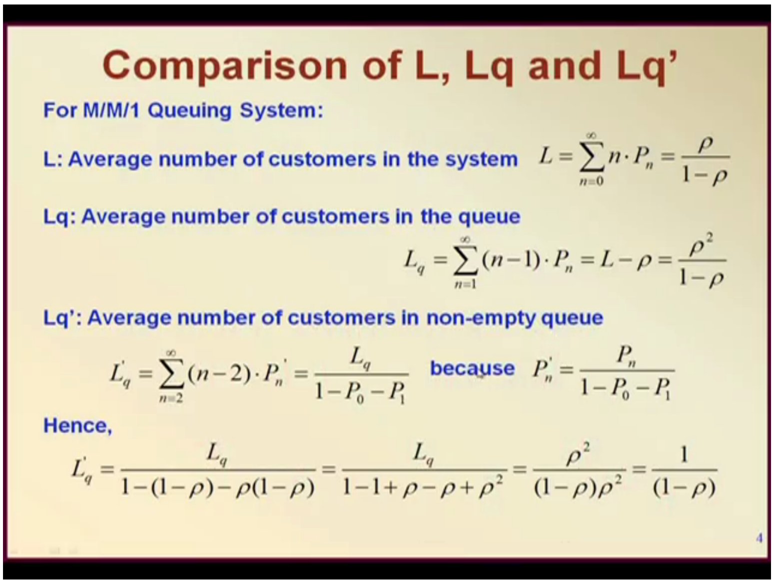
So look here only 1 minus P 0 minus P 1, what is 1 minus P 0 minus P 1? Let us work it out also, 1 minus P 0 minus P 1 equal to 1 minus 1 minus rho minus rho into 1 minus rho, because look at this formula. So that would 1 minus 1 plus rho minus rho plus rho square is nothing but rho square, right. So if suppose rho is half then rho square is 1 by 4. That means if the busy period is 50% then only 25% time there will be somebody in the queue, at other point of their will be nobody in the queue, is it alright?

So basically what we need to do we have to wait for that and LQ I mean LQ dash should be n equal to 2 to infinity, right, because 0 and 1 should be excluded and this should be n minus 2 into Pn dash, why Pn dash? Because the probability is are modified based on because Pn dash is like a conditional probability, let us write it here Pn dash equal to Pn by 1 minus P 0 minus P 1, so it is given the P 0 P 1 minus P 0 and P 1, but what is that value? This value is nothing but Pn by rho square, right.

So what really happens? That given that somebody is in the queue so that is Pn dash, so that is the probability that should be used really for calculating the expected number in the non-empty queue. So what you really cater here is the LQ dash. But look here 1 more interesting thing if n equal to 1 then this term become 0 once again, so that has got no value. So therefore n equal to 2, even if you submit n equal 1 you know it does not really make much of a difference and this term essentially then become is equal to the LQ by rho square. Because already we have found out that Pn dash equal to rho square, is it alright?

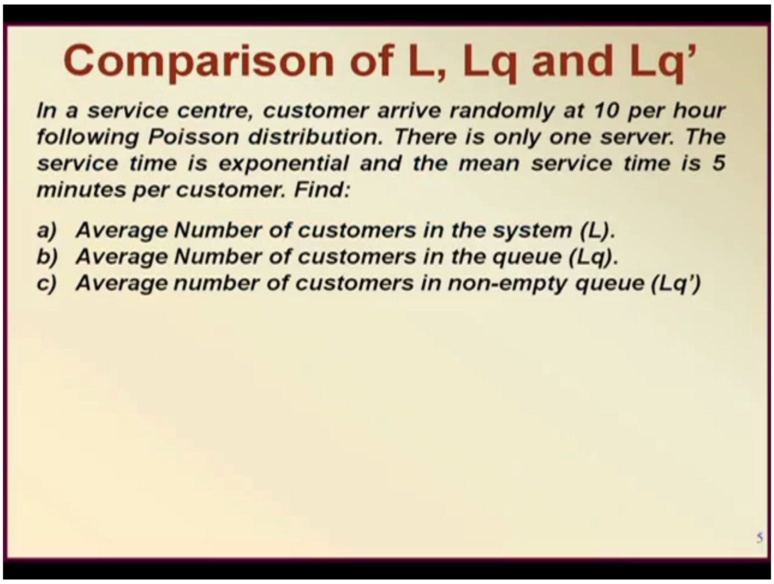
So the term essentially becomes the same and as the terms becomes the same you know we find that LQ dash formula comes out to be LQ by rho square, right. And what is LQ by square? LQ by rho square equal to rho square by rho square 1 minus rho because already this result is known, so that gives us LQ dash equal to 1 by 1 minus rho, right so this is the 3rd result.

(Refer Slide Time: 16:08)



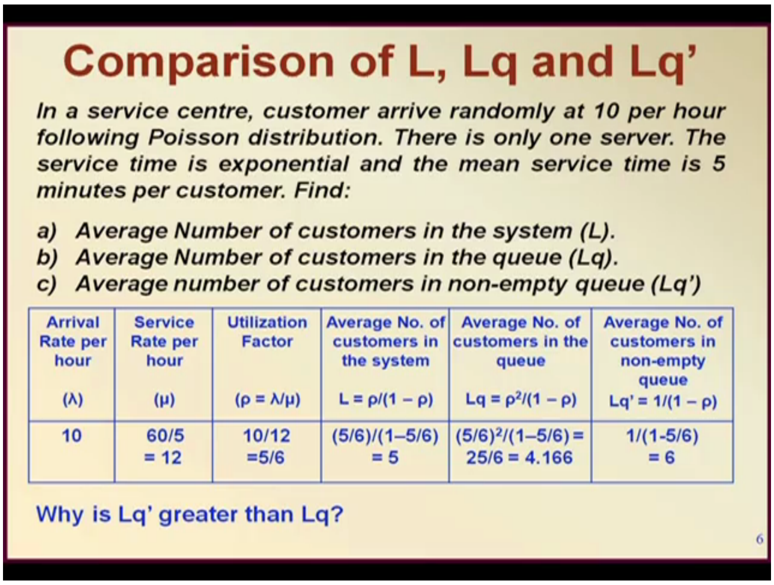
So these are the 3 results which are so important in this particular thing look at the slide once again L equal to rho by 1 minus rho, LQ equal to rho square by 1 minus and LQ dash is 1 by 1 minus rho, right. So this 3 results are we shall use and let us look at their comparison in the form of the problem.

(Refer Slide Time: 16:28)



See in a service center, customers arrives randomly 10 per hour following poisons distribution. There is only 1 server. The service time is exp1ntial and the mean service time is 5 minutes per customer. So you have to find the average number of customer in the system L, average number of customers in the queue LQ and average number of customers in the non-empty queue. It is a very straightforward problem all you have to do is find the traffic intensity that is rho, right. Find the rho and once rho is found you can find out all the 3 parameters very easily.

(Refer Slide Time: 17:13)



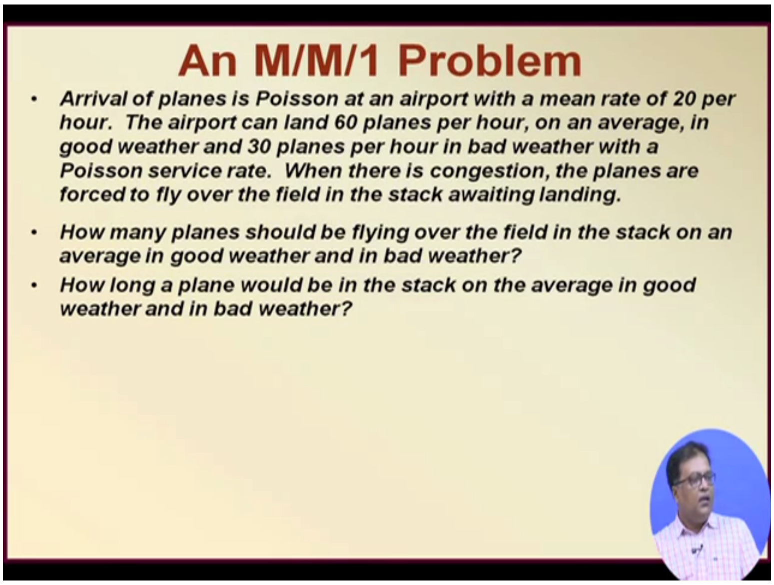
So here is the computation look here the arrival rate is 10, the service rate is 60 by 5 equal to 12, the utilization factor rho equal to lambda by Mu comes out to be fired by 6, average number of customers in the system is L equal to rho by 1 minus rho which comes out to be 5 and average number of customers in the queue LQ rho square by 1 minus rho comes out to be 5 by 6 whole square by 1 minus 5 by 6, so 25 by 6 are 4$166. And look what is LQ dash? LQ dash is 1 by 1 minus rho comes out to be 6.

So one important question, why is LQ dash greater than LQ? LQ dash 4$166 and LQ is 6. Why LQ dash that is 6 is greater than LQ? Unless it can be understood in this way, you see suppose queue is formed in this particular case the rho value is 5 by 6, that means the utilization is 5 by 6. What is non-utilization? 1 by 6, is it not? Again 5 by 6 that means the queue is formed only $5/6^{th}$ of the time, but what is rho square? Rho square is 25 by 36, right, so 25 by 36 is the actual time out of 1, right, where there is some queue.

So when you are averaging the queue over a smaller period of time rather than the full period obviously you get a higher value, right. So basically while you average you take out those data queue length is 0 if you take out those data then the remaining queue will become larger. So LQ is 4$166 and LQ dash is 6, that is why the LQ dash is larger.

Now what is the important of this? Sometimes you see we kind to average things only what we see. Suppose in a particular counter you know there are people there are no people, if there is none there is nobody in the system, obviously we may think that there is no queue letters only take data when there is some queue and then average it out that figure is LQ dash right. So one should be very clear about these 3 parameters L, LQ and LQ dash.

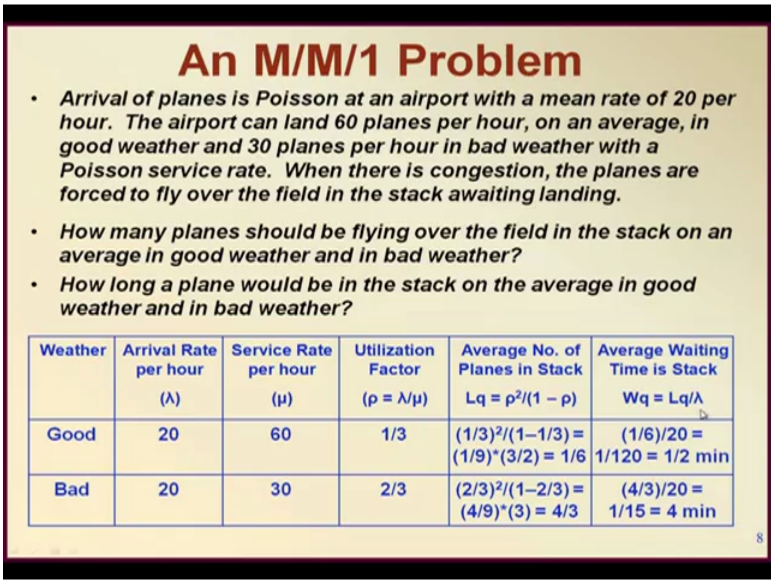
(Refer Slide Time: 20:08)



Let us look at some other problem, now this is an interesting problem about the arrival of planes. Arrival of planes at an airport is the mean rate of 20 per hour. The airport can land 60 planes per hour on an average in good weather and to 30 planes per hour in bad weather, obviously all this data that we are talking about in a peak period.

So the question is that when there is congestion the planes are forced to fly over a field in the stack awaiting landing, right. So there is a stack area, the runways is busy, the other planes that arrived has to wait, right. Wait where? In the sky, right, it has to move around and there will be awaiting landing.

(Refer Slide Time: 21:23)



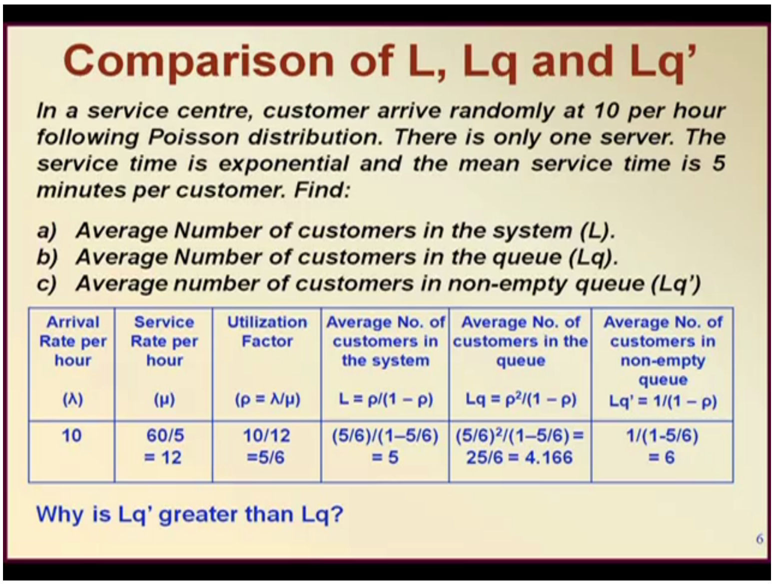
So the question is how many planes should be flying over the field in the stack on an average in good weather and in bad weather, right? And how long a plane would be in the stack on the average in good weather and in bad weather? Now the question now look at this diagram, now tell me which one is stack in this diagram and which one is the runway and which one is the queuing system, Is it alright? See the plane are arriving, the service begins when it is starting to land that means when it is getting the runway and actually lands.

So the service mechanism consist of the runway, that you consist the stack is actually the queue. So in this case we are really talking about the queue and not talking about the runway or the service mechanism, that means we are basically asking for LQ and WQ parameters. This point has to be very much should be very clear sometimes the service is more important and some other time the system is more important, sometimes the queue is more important, is it not?

Suppose you are waiting to get into a movie hall. When you are trying to get into a movie hall are hall you are gone there to see a movie, so seeing the movie is not you know something that is you know you are really worried about because you actually has gone for that, right. It is the queue outside that is what you are not liking, but on the other hand suppose you have gone to get a tool and you are waiting in the queue right. How long you waited in the queue is obviously important, but actually what is more important is how much time actually you spend in the entire process, right.

You spend some time waiting in the queue, you got some time you have spent getting the tool and then you come back. So therefore the entire time in this case the system times that is system that is W that is the waiting time in the system is more important, so this must be understood in the appropriate context.

(Refer Slide Time: 23:32)



So again look here in this case in this particular example you know this is the plane example, in the good weather the arrival rate is 20 and the service rate is 60, whereas in bad weather arrival rate is 20 and the service rate is 30. So what is the utilization factor? In this case rho is 1 by 3 and in this case rho is 2 by 3.

So if you compute LQ, the LQ is 1/3rd square by 1 minus 1 by 3 comes out to be 1 by 6, right. And in the bad weather it comes out to be 2 by 3 whole square by 1 minus 2 by 3 comes to be 4 by 3. That is the average number of planes in the stack that is LQ and if you divide this by lambda but please remember the Little's formula basically tells LQ by lambda bar.

The lambda bar essentially means if the lambda value is different over the period then we must take an average and divide by that and not simply lambda but most of the example we are taking the lambda is same all the time, so lambda and lambda bar are equal, is it alright? So that is why when we use Little's formula we are just dividing by lambda.

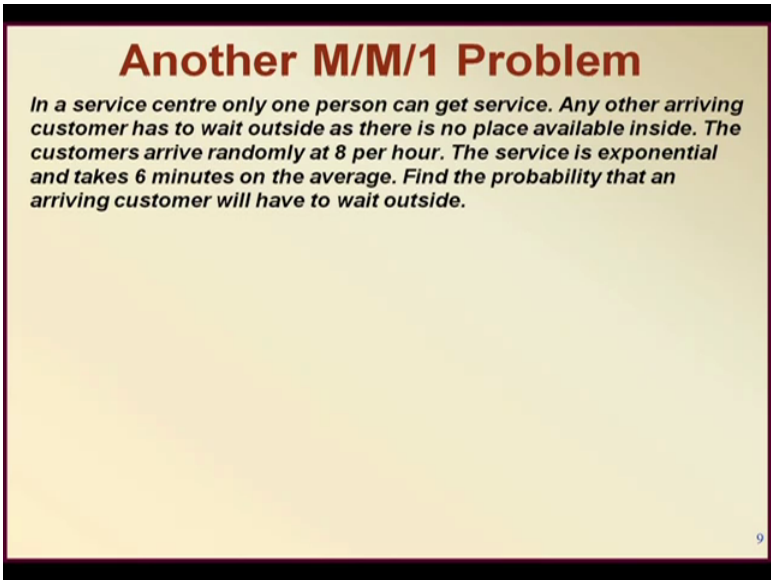
So in this case what we find the average waiting time in stack becomes WQ equal to LQ by lambda and which comes out to be half minute, right for the good weather and the bad weather it becomes as high as 4 minutes. So you see the waiting time increases 8 times, but you see sometimes we tend to think that you know that airport can land 60 planes per hour where is it is now doing only 30 planes per hour, so how much how this waiting times has increased by 8 times.

The service has fallen by 2 times but you know if you simply think it would be linear and waiting time in the queue will also be double that is not true, it is gone as high as 8 times. The reason is that even in good weather or the bad weather, even look at the bad weather the utilization factor is 2 by 3, that means there is 33 percent of time that is 1/3rd of the time you know the system is empty, there is no plane, is it alright?

So it is not that you know the all the time the planes are waiting, is it not? Planes are waiting for certain period of time and when those times are accounted for then this is happening. So these points must be remembered and please remember again and again I focused on this particular point that the mean value is 20 per hour that is the arrival rate. The service rate has to be higher than the arrival rate, only then the queuing problems will come to be an equilibrium and this kind of analysis that we are doing is meaningful.

But if the arrival rate is higher than the service then there is no equilibrium, if there is no equilibrium the system is in transient stage and in transient stage none of this formula is going to work, right. So that means rho should be always less than 1 that is the very important requirement.

(Refer Slide Time: 27:12)



Now let us look at another problem that in a service center only 1 person can get service. Any other arrival customer has to wait outside as there is no place available inside. The customer arrive randomly at 8 per hour, the service is exponential and takes 6 minutes on the average. Find the probability that an arriving customer will have to wait outside, right. So we have can done one problem earlier where we have seen that what really happens if you know it is when there is not enough facility available for the waiting customers.

Please remember it is not limited queue, we are not saying a customer who arrive will go away if there is no seat. It simply says that the customers will not be able to get a seat. Now this situation is different from another situation where the person if the system is busy will go away and that process is called is balking, is it alright? Anyhow think about this problem, stop here and will continue from here in our next class, thank you.