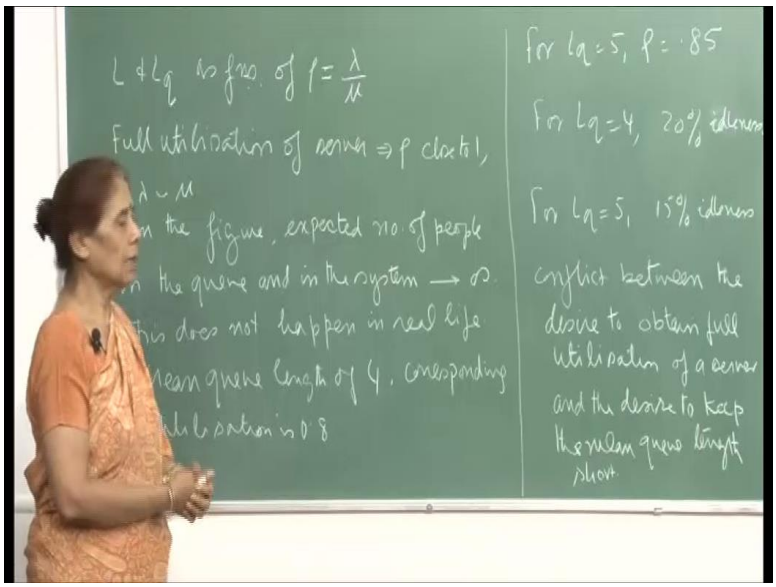
Introduction to Probability Theory and its Applications

Prof. Prabha Sharma
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture - 35
Analysis of L Lq W and Wq M/M/S Model

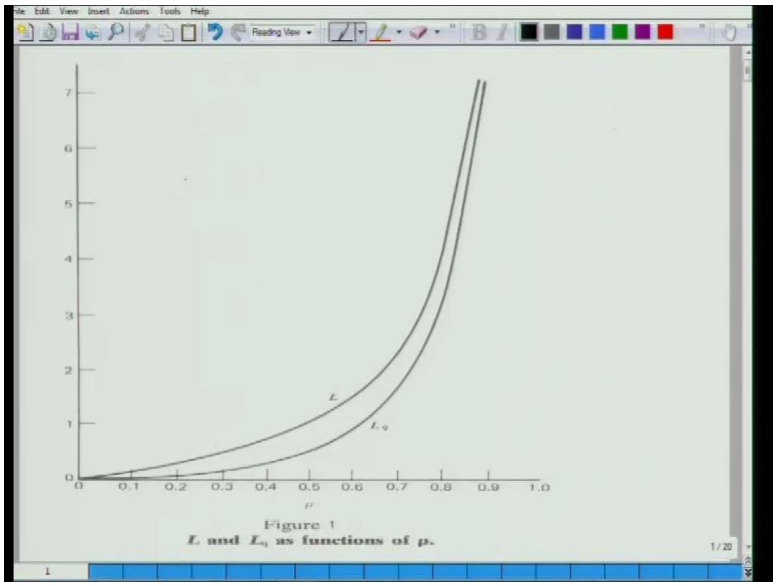
(Refer Slide Time: 00:15)



So, I will continue my discussion about the you know the queuing system. You can say the features that are important in determining various aspects of a queuing system. So, L denoted our average number of people in the system; L q was the average number of people in the queue. And then we also had computed the average waiting time that a person will spend in that system which includes waiting in the queue plus service time.

And then we will also talk about the waiting time in the queue which we show also important. So, anyway, so let me just continue the discussion about L and L q. And so this is as functions of rho which is your lambda by mu, that is your lambda is mean arrival rate and mu is the mean service rate. So, lambda by mu denotes your innocence utilization of the queuing system.

(Refer Slide Time: 01:27)



So, now we can see from the figure that is on the screen that as you know full utilization of the server; that means if rho is close to 1 then you see that your graph for L and L q both are going to infinity, right. So, that means, when lambda is close to mu, if rho is close to 1 that means, lambda is close to mu that is the mean arrival rate and the mean service rate are almost the same.

In that case you see the number of people in the system and the number of people in the queue they will the expected number will go to infinity, right. You can see the vertical, the approaching, the both the curves are approaching the vertical line. So, now this what we are saying is mean values; that means L is the average number of people in the system and L q is the average number of people in the queue which are, which going to infinity.

So, it is not, see, it is bad if once in a while your system has infinite people or very large people or the queue has becomes very large. But here it saying that it is the mean behavior that is on the average system will have infinite people, very large people, and the queue will also be very large, of course, ok. So, that is not acceptable, finely.

But, anyway, also what we want to say here is that now in this case this does not happen in real life because if the big crowd then it turns away lot of people. So, it is not that the queue continues to grow. So, that does not really happen. And therefore, what we are saying is that in real life there will not be a balance because the queue system, you know, how many people turn away, and so on.

So, it will not follow the same pattern; and therefore, when things become so bad we cannot apply the same rules that we started with, you know, our assumptions. And therefore, we will say that the system is not in balance. So, therefore, it cannot be, you know, measured; it cannot be analyzed by any of these models that we have written down, right.

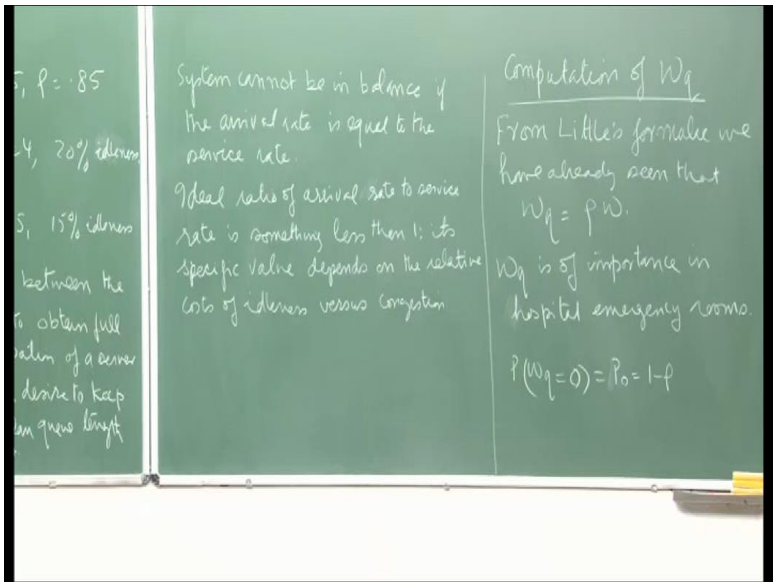
So, if you look at the mean length of 4 on the vertical line or then draw horizontal line from 4 then you see that it will meet the corresponding utilization is 0.8, ok. So, mean queue like that; that means, on the L q curve. So, and so L q is the first curve and then L is the second curve. And similarly, if you look at the value of L q equal to 5, the corresponding rho value is 0.85; so that means, when you allow 4 people to wait in the queue then the server will be ideal for 20 percent of the time, right; because rho is 0.8.

So, 0.2 is the fraction of time that the server will be ideal; and so 20 percent of the time the server is ideal. And if you allow the queue size to be 5 then it will be 15 percent idealness. So, you can see that there is a definite conflict that is if you want that your server should not be ideal for very long time or even not then as we saw that the, if the system is allowed to operate freely. Then your queue and the people in the system and people in the queue will become very large.

So, therefore, but then you do not want that to happen because then you lose out the customer, and so on; and it becomes a chois. So, there is a definite conflict between the issue, between the desire to obtain full utilization of a server and the desire to keep the mean queue length short. So, you can see that. So, where, wherever there is good will of the customer is very important. Certainly, the persons who are offering the service would want to make sure that the queues do not become too big.

But then there it is more important, that means, the financiers are important and you are having a service where the, you cannot keep too many servers because that means, that many salaries, and so on. And then of course, you are, the servers will be ideal for a long time. So, therefore, you, one has to balance. So, the system can; of course, first of all this model has shown you that the system cannot being balance if the arrival rate is equal to the service rate, or even if the arrival rate is close to the service rate.

(Refer Slide Time: 06:03)



And so the ideal ratio of arrival rate to service rate is something less than 1; that is acceptable because your service rate must be more than the arrival rate, otherwise and that make sense. And so but its specific value depends on the relative cause of idealness verses congestion. So, if, for you it is very important that the system should not be congested; there should not be too many people in the system. Then you make sure that your service rate is much higher than the arrival rate.

And, if your, you have limitations; you may not to provide, you know, more than 1 server, may be let us say or is efficiency of service providing the service, then you know, we will go for; so I mean, right, so idealness verses congestion. So, if you want the, allow the server to be ideal you will not have that many people waiting in the system, right. Because, your service system will be such that your mu is much higher than the arrival rate, you will ensure that.

And, in that case, your, L and value of L q will be small, but if your mu is close to your lambda then there will be congestion. So, one has to really strike a balance between, you know, idealness verses congestion. So, this is the whole idea I mean. So, therefore, you see we can, through these, this model we can study and decide what should be our level of service, and given what is your level of, you know, customers arriving to the system, right, ok.

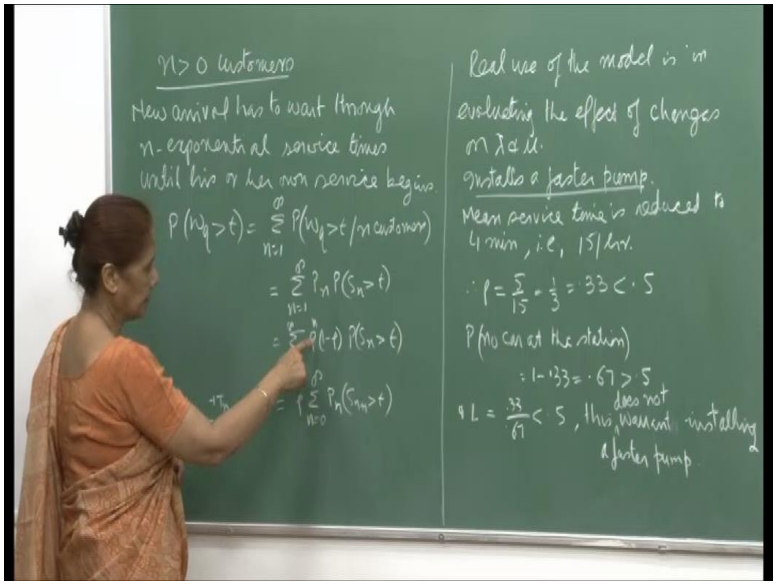
So, now, even though we, the little's formula told us that if you, we have already computed average waiting time in the system. And now we also know that the W q can

be written as rho times W; that means, the average number of people waiting in the queue is utilization factor times the average waiting time in the system. But, we will also like to compute this independently because the distribution of W q is also of importance. And besides, the W q is of importance in hospital emergency rooms.

Because, in a hospital emergency room the time that you are waiting to be, you know, taken to the doctor is important because you want to cut that shot; it is an emergency room and therefore, people need treatment fast. So, you would, therefore, W q is of importance; and in hospital emergency rooms you would want to ensure that your W q is not very large. And so we want to look at it in a greater detail and also obtain the distribution of W q.

So, now, of course, you can immediately write down this relationship because if there is nobody in the system when the arrival comes for emergency treatment then probability, so probability of the waiting time will be 0 because the patient will immediately be taken to a doctor for being treated. So, there in that case your waiting time will be 0. So, probability the W q is 0 is P 0; that means, there is nobody in the system which is 1 minus rho. So, now we will want to compute the distribution when person coming to the system has to wait; that means, there is 1 row more than 1 person already in the emergency room and then the person has to wait. So, this we will compute; try to obtain the distribution for W q also.

(Refer Slide Time: 09:49)



So, for n greater than 0; that means, if the any customer is already present in the system then the new arrival has to wait through n exponential service times until his or her own service begins, right. So, if you want to find out the probability, the W q is greater than t, where W q is the waiting time in the queue. So, then probability W q greater than t, so I will break it up into conditional probabilities and then add up.

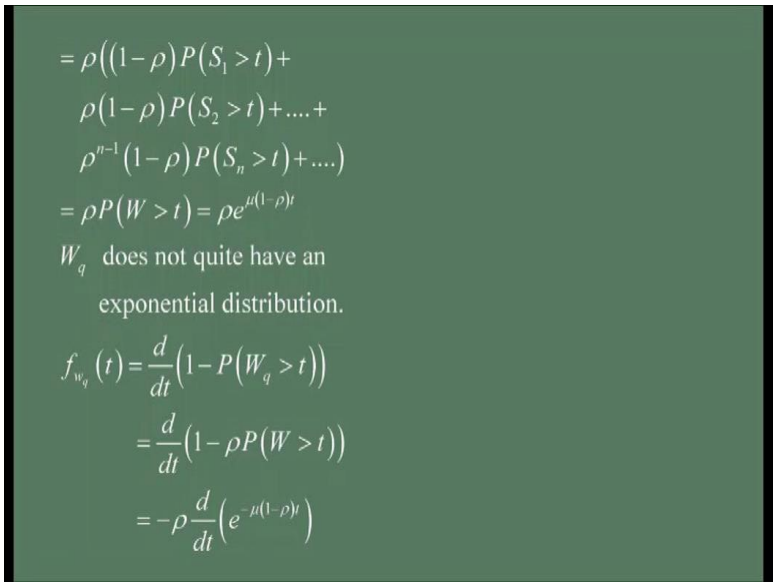
So, this is will be probability W q greater than t given that there are n customers, and so we add up from 1 to infinity because there has to be atleast 1 customer to being serviced then only there will be a need to wait in the queue. So, this is n to 1 to infinity probability W q greater than t given that there are n customers in the system. And this then would be written as; so now, when you write this probability then this will be P n into probability S n greater than t because if these are the n service times then S n would be t 1 plus t 2 plus t n. So, this has to be greater than t because the n people being serviced.

See, you are in the, you are waiting in the queue, so then n services have to be completed and that takes because your waiting time is more than t; that means, these services are taking more than time t; and this into probability of there being n customers. So, then this will be; so I will substitute, I will write down for P n which is rho n into 1 minus rho, n varying from 1 to infinity, probability S n greater than t, right.

Now, if I take rho outside and then I write this as rho, so this will be a rho n minus 1 into 1 minus rho, and this is probability S n greater than t. Now, see, at n equal to 1 the value here is rho into 1 minus rho, and this is probability S 1 greater than t, so if I rewrite this as rho outside, and now at n equal to 0, your p n would be just 1 minus rho because rho n, rho raise to 0 would be 1. So, this would be 1 minus rho. So, same thing; and here it will be S 1 greater than t. So, the same event is being written here and therefore, all subsequent ones will be the same.

So, this helps you to, because now you have the summation; probability sign is missing here. So, please write; this will be p in, probability of S n plus 1 greater than t; I will just write it; there should have been probability; so probability S n plus 1 greater than t. So, therefore, so now, you have, because the summation is from 0 to infinity, so then we can sum it up easily.

(Refer Slide Time: 12:42)

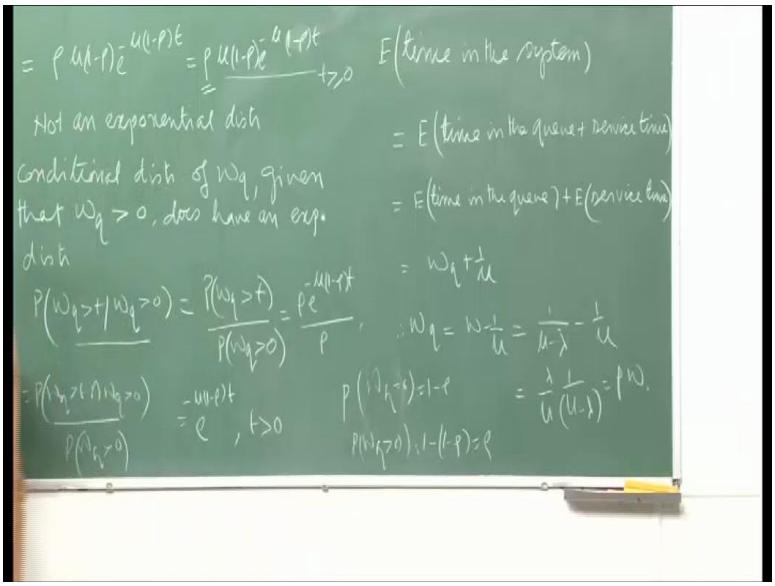


And therefore, you can just expand this. And so here this is no people then the, service time of the, sorry; 1, rho into 1 minus rho actually would be the, you know, 1 person in the system; then the waiting time is the service time of the first person is of the 1 person already present in the system is more than t, and so on. So, when you take out rho then this is what, we have, you just expanding this expression like this, this series.

And this, now we know; we recognize this because when we computed the distribution for the average time in the system. So, then this was, this is nothing but, and that is why we wanted the sum from 0 to infinity. So, this is probability W greater than t. And we already know the distribution of W greater than t; so we know this probability. So, therefore, this rho into e raise to minus mu, 1 minus rho t. So, we have computed this probability W q greater than t which is equal to this.

Now, so the rho is creating the problem because I cannot call this an exponential distribution, why? Because, if you write the p d f of W q, then this will be, you know, this you will write as 1 minus probability W q greater than t, yeah; this whole thing; so then d d t of that will give you the p d f of W q, right. So, this will be minus rho into, because this is 0; and I will write down the, yeah; so this one is rho into e raise to minus mu, 1 minus rho t; substitute for this here, then d d t of this, right; rho is outside.

(Refer Slide Time: 14:31)



So, then differentiation gives me this expression which is equal to this. So, therefore, this would not be an exponential distribution because exponential distribution will be this; rho is the extra part, right. But, if you consider the conditional distribution, so now, that is important; conditional distribution of W q, given the W q is greater than 0, does have an exponential distribution because now you will write the conditional part.

And, this will that be equal to probability W q greater than t because when you take the intersection, obviously, t is positive; so W q greater than t. So, the intersection of these 2 becomes just this event, and divided by probability W q greater than 0. See, late this you have to write as probability W q greater than t intersection, W q greater than 0; and then it will be probability W q greater than 0; this is what our formula for conditional probability is; but, this is the same as this; of course, given that t is positive.

So, this event is equivalent to this event, given that t is positive, right; and then you divide the probability W q greater than 0. So, therefore, this becomes now, rho e raise to minus mu 1 minus rho t divided by rho, because probability W q greater than 0 is; this is what we computed here; W q is not 0, yeah. So, remember our W q, probability W q equal to 0 was 1 minus rho, right.

Because if the person does not have to wait that there is no person in the system, and so therefore, P 0; this probability is equal to P 0, which is 1 minus rho. So, if you want probability W q greater than 0 then it will be 1 minus of 1 minus rho which is equal to rho, right. So, probability W q greater than 0 is rho; so we divide this. Now, the rho

cancels out; and this is e raise to minus mu 1 minus rho t. So, this now is coming from; that means, this now represents because the conditional part W q greater than t given that W q is positive. So, this is this which matches with our exponential distribution; and so the conditional distribution of W q is exponential with parameter mu into 1 minus rho t.

So, this is therefore, I again just want to repeat that this is not conditional distribution; we just removed this. So, it is just probability W q greater than t is that what you are computing; you started writing out of this; and then this is, I will just; no, we wrote this expression and just to, so that I can relate this series with this series with probability w greater than t; so that was helpful.

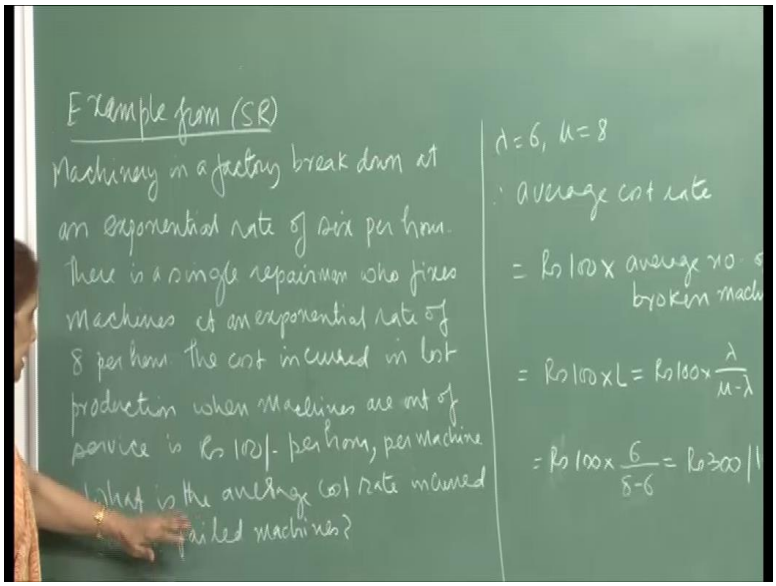
And then we saw that the distribution is not matching with the exponential. So, it is only when you take the conditional probability that you will get, right. And so this already we had seen; and therefore, this is rho. So, that is it. So, therefore, and this will be useful at times you may really want to know the distribution of the; so this will be the conditional distribution of W q which you can recognize.

Otherwise the distribution of W q is given by this; I mean the f q t, f W q t, right. Now, let us compute the expected time or the average time spent in the queue waiting for to be serviced; so that we will now compute independently because this is only the conditional part. So, you will say that the expected time in the system is equal to the expected; so the time in the system is time in the queue plus service time; so expected value of that.

So, expected time in the system is expected value of time in the queue plus, the service time right. And so this can be rewritten as expected time of time in the queue plus, expected service time. So, this if we denote by W q. So, this is the convention way we have been following that the variable is also treated as the, S denoting the expected value of that variable.

So, this is expected time in the queue plus, expected service time is 1 by mu, right; because your number of services is with parameter mu, exponential with parameter mu. So, 1 by mu is the expected service time. So, this is what you have; and therefore, W q is W minus 1 by mu; and you can, so we know W, expected value of W which is 1 upon mu minus lambda minus, 1 by mu. So, this becomes what we had computed, rho times W. So, W q is equal to rho times W, and which we have already seen through using the little's formula.

(Refer Slide Time: 17:21)



So, we saw that the conditional distribution of W q and W are both same exponential, and the parameter was, yeah; it means, I mean the same parameter. So, they, now also want to make an observation that the little's formula that we obtained under special conditions, but it turns out. So, that means, the relationship between L and L q, W and W q, and L and W, so they are all; in fact, what it means is that if, you can find any one of the quantities, L, L q, W, W q, then you can find all the other 3.

So, the all the 4 are related, and it turns out, fortunately that under very general conditions this formulae are valid. So, therefore, you know, computing any one of them would help you to get the values of the others. Now, I want to continue this discussion on these quantities L, L q; also show you how we then through these analyze these waiting systems have been queuing model.

So, this particular example; it is the small one; but, anyway this is from Shelton laws; and what it says that machinery in a factory break down at an exponential rate of 6 per hour, ok. So, that means, the arrival of the machinery for repair is 6 per hour. And then there is single repair man who fixes machines at an exponential rate of 8 per hour. So, the arrival and the service rate as all provided to you.

The cost incurred in lost production when machines are out of service; see the machines come for repair; they are waiting; so the repair man has to repair then. So, while the machines are out of service they incur a cost because you, the lost production. So, rupees 100 per hour per machine is lost to the organization, right. So, the cost of lost production

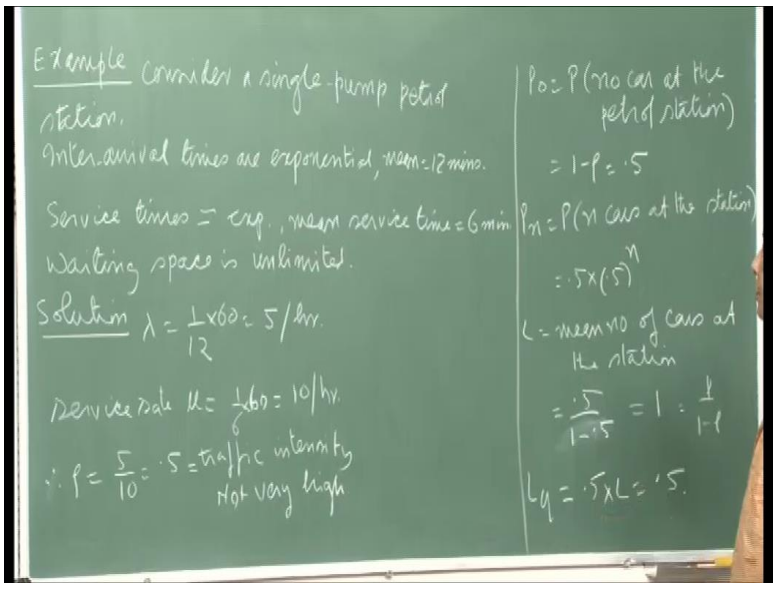
is rupees 100 per hour per machine; this is what is given to us. Now, we want to find out what is the average cost rate incurred due to failed machines.

So, while this is, the machines are waiting to be repaired, they are not producing, and therefore, there is a loss to the organization; and this is the rate. So, you want to find out the average cost rate. Now, the average cost rate will be dependent on the number of machines which are in the system; which are either waiting to be repaired or which are being repaired. So, that will be our number l.

So, therefore, you see, L is lambda is 6, mu is 8. So, therefore, average cost rate we will write as 100, rupees 100 into, average number of broken machines which is either waiting to be repaired or they are waiting, or they are being repaired. So, this will be rupees 100 into L; L is the average number of broken machines which are in the system; L gives you the average number of people or customers in the system.

So, therefore, this is rupees 100 into lambda upon mu minus lambda, which is 100 into 6 upon 8 minus 6; so rupees 300 per hour; so the loss. So, therefore, this is the very important parameter because the system would now like to evaluate whether the repair man that they have is good enough, or they need to have more repair man; because, it depends on what is the, how much the loss to the system compared to the salary of a repair man, and so on. So, that question you see will always be running through all these examples; and you want to analyze; and you know, this should hopefully help you in your decision process.

(Refer Slide Time: 23:49)



Let us take another example. So, you have now a pump station, a single pump station; so single pump petrol station. So, there is only 1 pump, and so cars that come for taking the petrol have to wait in the queue. So, 1 is being serviced; once the serviced car is done with then the other will come from which are waiting in the queue.

Now, inter arrival times are exponential with mean 12 minutes; and that means, inter arrival time, the average time between 2 arrivals is 12 minutes; and the service time is exponential again with mean service time 6 minutes. So, average time it takes to fill up the car is 6 minutes; and waiting space is unlimited. So, I am not put any question here, but as we go long we will see what are the kind of questions we want to answer here.

So, let us see; lambda therefore, that means, the arrival rate is 5 per hour because arrival time, inter arrival time with mean is 12. So, therefore, the number of arrivals per hour is 5. And similarly, the service rate mu will be 1 by 6 into 60 because this is per minute; this is only in minutes. So, you convert to hours. So, this will be 10 per hour; that means, the service rate you can, on the average you can fill up 10 cars in an hour, right.

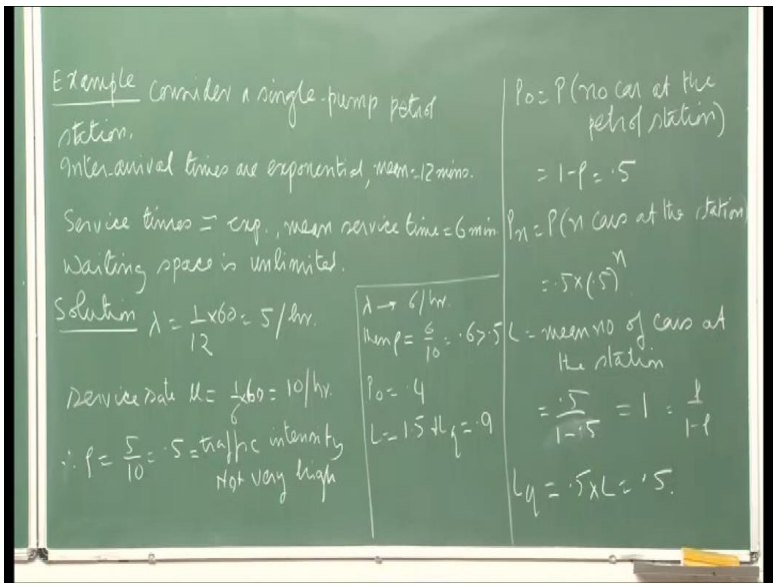
Rho the utilization factor or the traffic intensity, we have lot of names for this. So, rho is 5 by 10, right; 5 is your arrival rate and 10 is your service; so 5 by 10. And therefore, this is 0.5. So, the traffic intensity is not very high; 0.5 is not considered to be very; that means the petrol station is not very busy right now. So, the kind of service rate and the kind of arriving rate.

Then, if you want to look at the probability that there is no car at the petrol station; so that will be 0.5 1 minus rho which is 0.5 again which is the high probability. Then, P n is the n cars at the station; so that will be 0.5 into 0.5 raise to n, right. And then the mean number of cars at the station, mean number of cars would be lambda upon mu minus lambda; or, well, this is, why I am writing this as, how, in fact, this is, sorry; this is rho upon 1 minus rho. So, the mean number of cars at the station is 0.5 upon 1 minus 0.5 which is 1; and L q, the car waiting in the queue is 0.5. So, this is the idea.

So, now, we want to analyze the system through these quantities that we have computed. And you know, like you want to again answer the question that in case the arrival rate increases then supposedly the traffic intensity will go up, because this number will go up. And then what kind of numbers L and L q will be there; the values will also go up; and therefore, the petrol owner, the station owner may want to ask a question is, should we install a faster pump and so on.

Or, of course, more than 1 pump you will, the system that model we will discuss later on when you have more than 1 server; right. Now we are talking about only 1 server systems; and therefore, the only option that the man may have, in case the arrival rate goes up, option would be to install a pump which is filling up cars are at higher rate. So, we will just look at the analyses with the numbers.

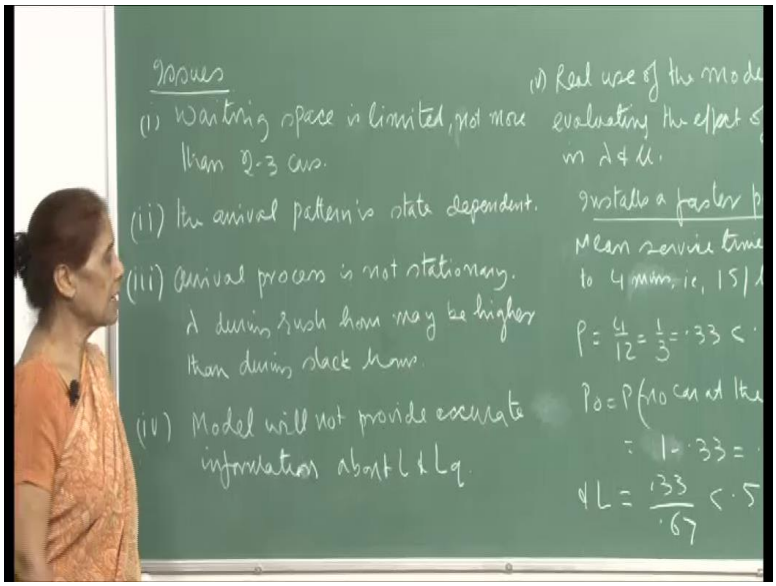
(Refer Slide Time: 27:44)



So, through this examples which we are just discussing, I again want to raise the issues about validity of a model. So, it is very important that you keep recalling what are the assumptions under which we are working, and what are the; so for a example here I wrote that waiting space is unlimited, but you know that in a petrol station waiting space cannot be unlimited; and usually when you have space for 2 to 3 cars.

So, but of course, the, with the given data right now, it did not really matter because your L was 1, and your L q was quite small, right; L q was 0.5, I think; L q is 0.5. So, therefore, that, right now it is not an issue whether the space is limited or unlimited, but in case your data changes then to say that your waiting space is unlimited, it is not a very valid assumption.

(Refer Slide Time: 28:39)



So, therefore, one should always keep this in mind that the better model would be when you talk of queues with limited capacity or with limited waiting time. So, that would be a better model for such an example. Then, the arrival pattern is state dependent; now, here to assume that lambda will remain the same all the time is not correct because if already 2 to 3 cars waiting, the person may want to go to the next petrol station. So, therefore, this is not.

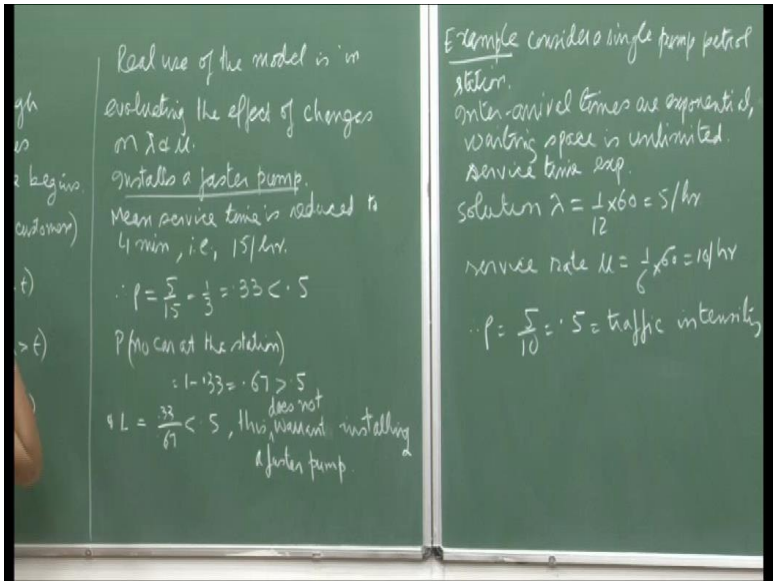
And, arrival process is not stationary also. So, it is state dependent and also not stationary because during the rush hour there may be, the lambda may be higher than corresponding to when it is lack hour. So, therefore, the lambda itself may change during the day. So, the lambda is not stationary; and it is also a state dependent; because people do not like to wait for too long, for because as you know you can always drive further and get another petrol station. There may be other considerations also; that is true. But sometimes if you like to wait at a particular station because they are familiar with the people know them, and there can be other, so many other reasons.

Then also we must keep this in mind that whatever computations we are doing, remember they are not giving us accurate information about L and L q, and the remaining other parameters W and W q. But, we can certainly change the values of lambda in mu of the system; so that means, we can study the changes in the system and then correspondingly see what are the changes in these numbers L, L q, W and W q.

So, this models certainly will help you to study the changes; whatever change is taken place in the system, then accordingly find out the changes in L and L q. So, that is what I want to; so that is what I have stated here, that the real use of the model is in evaluating the effect of changes in lambda and mu, right.

So, for example, the station owner has the choice of, has alternative of installing a faster pump. If you done that then the mean service time is reduced to 4 minutes per car; it was earlier 6 minutes per hour on the average. So, now then the average time has gone down to 4 minutes; that means, the petrol pump can service 15 cars per hour.

(Refer Slide Time: 31:21)



So, by installing a faster pump the mean service time is reduced to 4 minutes; that is the pump can fill up 15 cars per hour, right; and so your rho width now become 5 upon 15 because the arrival rate is 5 per hour. And so the intensity, traffic intensity or I use to called it the utilization of the petrol pump, and so on, so that is 5 by 15 which is 1 by 3; and this is 0.33. So, this is less than 0.5. So, the earlier one was 0.5 traffic intensity, now it has come down to 0.33.

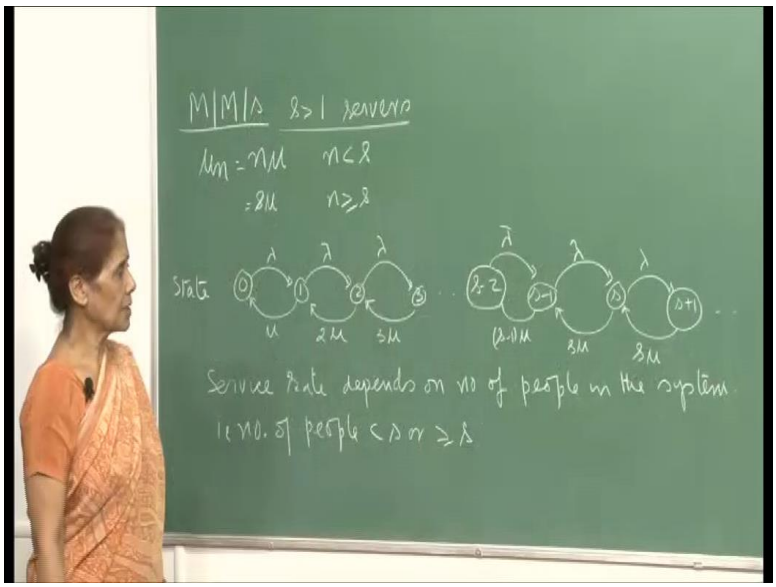
And, the probability that there is no car at the station, so that comes out to be 1 minus 0.33, right; because it will be 1 minus rho; and so that is 0.67 which is greater than 0.5; so that means, the petrol pump would be vacant for more time. The fraction of time would be higher than here because there it was 0.5; since rho is 0.5, so 1 minus rho is also 0.5.

And, your L, the average number of people in the system or at the petrol pump, that means, the number of cars getting filled up or waiting to be filled up, so that will come out to be 0.33 by 0.67 which is also less than 0.5. So, therefore, this does not warrant installing a faster pump because your petrol pump is vacant longer, the intensity, traffic intensity has come down, and so on, yes. So, if you are looking at from the view point of the petrol pump owner then certainly it does not wanted installing a faster pump.

In case, the arrival rate goes up, that means, you have 6 cars per hour instead of 5 cars per hour, and with the current pump that you, the man has, then this will be, rho will be 6 by 10. So, this is the 0.6; that means your traffic intensity will go up from 0.5 to 0.6. Your probability of there being no car in the system is 0.4; and your L is 1.5; and your L q is 0.9.

So, therefore, you see that is what I am saying that, now with the model you can play around; and for different values of lambda we can figure out what is, how these numbers are changing. And then if you can, as I said, you know, the losing the good will of the customer verses the cost of installing a faster pump and so on, what can be, the owner can study all those things through this model, right. And therefore, that the basic contribution of this model lies in being able to study the various changes that will take place in your, you know, you can call them parameters when you change your lambda and mu change.

(Refer Slide Time: 34:19)



So, now I will, after having discussed 1 server model we will now look at the situation when there is more than 1 server. So, this is that model would be M M S, that the same, the pattern, or arrival pattern and the service pattern are the same, right; you can say that, and therefore, and the number of servers is now more than 1. So, in that case you see as long as there are number of people is less than s, then your service rate will be.

Because, whoever is there is in the system, if the n people in the system, all will be serviced; and therefore, your service rate will become n times mu; that is understandable because everybody is being serviced, and therefore, the n people who are being serviced simultaneously. So, the service rate you can say has gone upto n mu, right.

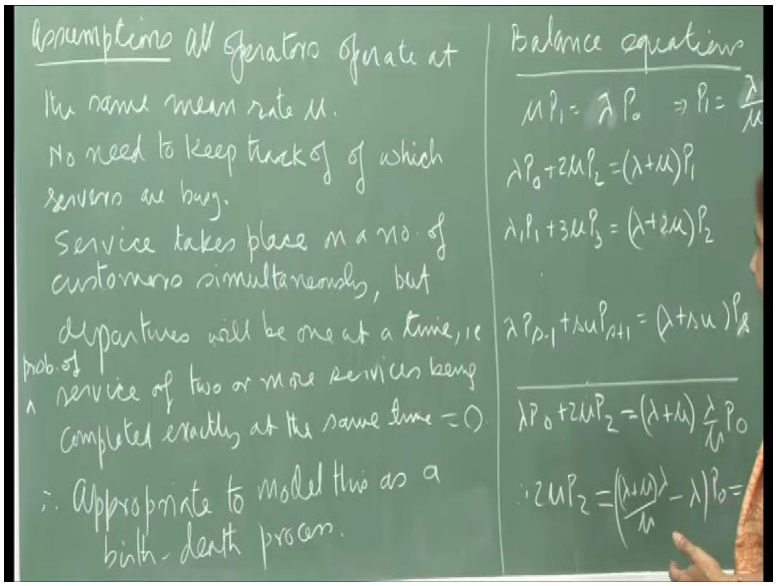
And, if you have more than n s people in the system, then of course, it will be remain at s mu because only s people can be serviced, since your s servers. And therefore, then the service rate will be s mu, but I will try to explain; and therefore, diagrammatically if you look at the transition diagram here, then you see arrival rate is the same lambda, right; which is, you know; so lambda is the arrival rate, but the service rate changes.

So, if you have 1 percent in the system then it will be mu, right; because there you serviced and therefore, you get back to 0 state. If you have 2 people then 2 are being service simultaneously and therefore, you are at the rate at which the system can transition from 2 to 1 will be 2 mu. So, this should be understood very well because so what we saying is that, that many people are being serviced and therefore, the probability of transition; that means, the rate at which you can transition from 2 to 1 will be 2 mu.

And, similarly with 3 people, the rate of transition will be 3 mu, but the arrival pattern is the same; and therefore, the arrival rate is lambda. So, this will go on upto s minus 2 and then upto s minus 1; and when you have s people in the system if your state is, the system is occupying state s, then it will be s mu. And thereafter, the service rate will remain at s mu, right.

The arrival pattern would be at the rate of lambda, but the service pattern, service rate would then remain at s mu. So, this is the idea. And so here this service rate depends on the number of people in the system; that is if the number of people is less than s than it is n mu, and if it is more than s then it is s mu.

(Refer Slide Time: 37:18)



Now, let us understand the assumptions; it is very important. And if the, what we are assuming is that all operators operate at the same mean rate. So, right now this is the simplification; because, obviously, it will get very complicated if I have different servers with different service rates. So, therefore, we are assuming that all operators operate at the same mean rate, mu.

Therefore, I am saying that the, when there are n people in the system, the service rate will be n mu; and when there are more than n, it will be s mu. So, this is possible only if I make this assumption that all operators or all service people are operating at the same mean rate mu, right. And so this makes it; because then we do not have to keep track of which servers are busy; you know, then we will have to accordingly keep changing the service rate; and that will become quite problematic, right.

Again, if I feel that this is not very, this can be treated as a realistic assumption because, you know, person may be a little more efficient than the other, but the differences cannot be very, very large to really take care of them in the system, right, in the model. So, this is what 1 assumption.

And then the second assumption which is important is that departures will be 1 at a time; that is probability of service of 2 or more services being completed exactly at the same time is 0. So, this probability, that means, at departures because that is the important of the, you know, when we say that we looking at M M S system, and yes I will have occasion to explain. So, then when we talk of Markov process we will discuss in detail.

And so anyway, I am, when we talking of Poisson process, remember I had told you that there is always a small, there is a small enough interval in which we will say that the probability of 1 arrival is probability of, we know, the is something like lambda delta t, right. And then for more than 1 arrival it was of the order delta t square, a higher order right.

So, therefore, we were neglecting that; so the probability was again, that means, we assume that exactly at the same time 2 or more customers will not leave the system. So, the service will not get completed exactly at the same time for more than 1 people. So, therefore, there will be a distinct interval between 2 departures. So, therefore, we can model this as a birth, death process, because our birth and death process the assumption when we talking of M M S.

So, then the basic assumption is that your arrivals and departures are distinct at distinct times; you cannot have more than 1 departure or 1 arrival at the same time. So, once we make that assumption; and the second assumption is that the operators are all operating at the same efficiency; so the mean rate is same. Then we can process this system as a birth, death process, right.

And, so now, I can write down the balance equations. So, this is the first one which is easy to understand, right; because you have, you already have 1 person then the rate at which it can depart is mu. So, mu P 1 must be lambda P naught. 1 person can arrive when you have 0 when you are in the 0 state; so then this and this gives you this, right.

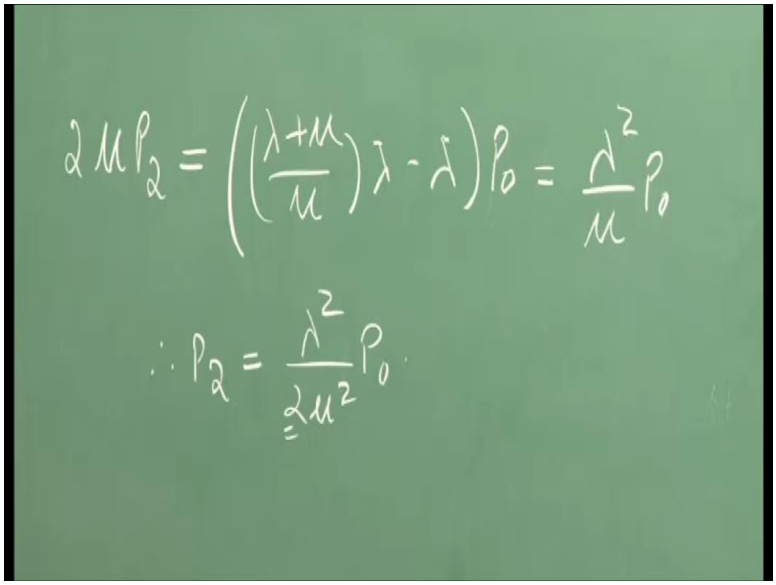
Similarly, when you, I want to write for, so 2 people in the system then it will be lambda P naught. You can go to P 1, sorry, to 1; and from here when you have 2 people then at the rate 2 mu you can go to again because 1 departure 1 person get serviced, and so you again to P 1. And here, it will be lambda plus mu, right, P 1; that is why the transition diagram. And even when I were discussing M M1 system, I had explained to you how you can, you know, interpret the transition diagram. So, it will be lambda plus mu.

So, it is actually not; there is nothing new here except that you have to remember that; so I have not written down the remaining things; it is understood that upto s minus 2 you will have this thing. And then after that you will, the movement you have s people in the system as a more; then this is the balanced equation that you will get, right.

So, once you have written down these balanced equations, immediately you can start solving. So, P 1 in terms of P naught will be lambda by mu P naught. And then if in the

second one if you substitute for P 1 in terms of P naught, then you get your 2 mu P 2 as, ok.

(Refer Slide Time: 42:03)



So, the expression 2 mu P 2 simplifies to, equal to lambda square upon mu P naught. And therefore, P 2 is equal to lambda square upon 2 mu square P naught. So, note the correction, that 2 was missing. So, it should be lambda square upon 2 mu square P naught. So, all the probabilities can be computed in terms of P naught.

So, important thing is that the moment you have more than 1 server things change a little and one has to understand, and what assumptions you are now modeling the situations; and so I try to explain to you how we will, under what assumptions we will treat this as a birth, death process. So, the service is, are all at the same rate; that means, all servers have the same efficiency, and that no, more than 1 departure at exactly at the same time. So, there will be just little interval of time between any 2 departures from the system. So, under this you can easily write.

And then of course, this service rate changes depending on the number of servers you have. And so and all these 3 assumptions you can write these balance equations, and then we will try to get the probabilities in general formula, and then we will again compute your quantities L, L q, W, W q to get an idea about. So, therefore, your traffic intensity also will change, right; you can see that; because your service rate is changing and therefore, your traffic intensity will also change. So, all these again open up a very interesting this thing, you know, situation; and we will like to look at them.