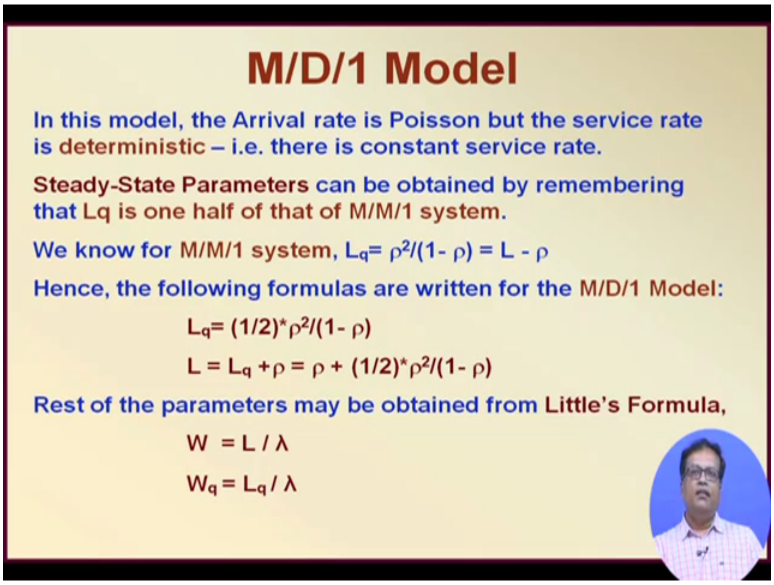
Decision Modelling
Professor Biswajit Mahanty
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur
Lecture 17
M/D/I and M/M/s Queuing Models

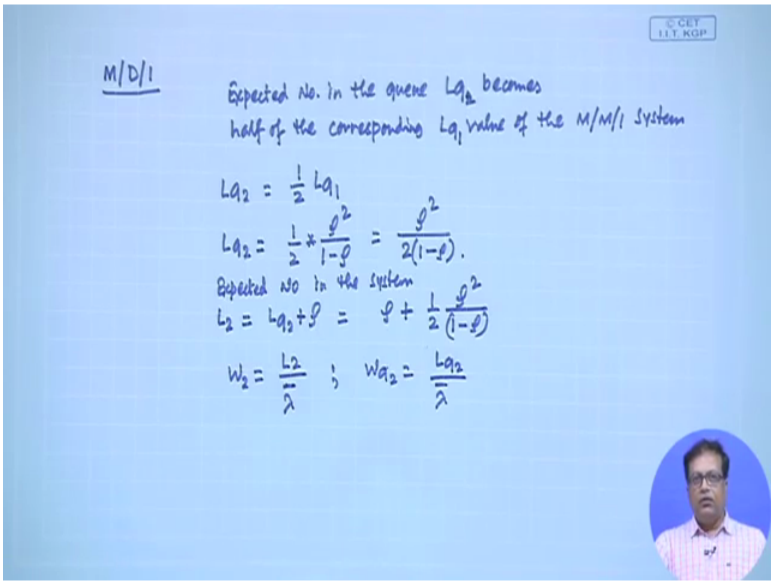
Right, today we are going to discuss 2 types of queuing models, specifically the M/D/1 and M/M/s queuing models. So far we have many seen the birth and death process and thereafter we have seen M/M/1 queuing models and various examples of M/M/1 of M/M/1 queuing models, various other issues but let us now go ahead with variations in them and 2 very important systems which we really look into, one is the M/D/1 model and the other one is M/M/s queuing models when there are multiple servers right. M/D/1 is still single server, but it is a deterministic model the service is deterministic. In M/D/1 model what happens the arrival rate is Poisson, but the service rate is deterministic.

(Refer Slide Time: 1:24)



What is the difference? The service rate is not probabilistic but it has got a fixed value sometimes what happens the service rate is going to be exactly the same suppose it is a computerised service system and as long as you give your input correctly and things are more or less in order then all the service time will be exactly equal right, will it make a difference that is what we are going to examine that is supposing the service rate is not probabilistic, becomes deterministic then what exactly happens under such situation. We are not going into details of calculation but we shall we re-evaluate steady-state parameter values, particularly the 4 parameters that we know; one is the number expected number in the system that is L, expected number in the queue that is L q, the waiting time in the system W and waiting time in the queue that is W q.

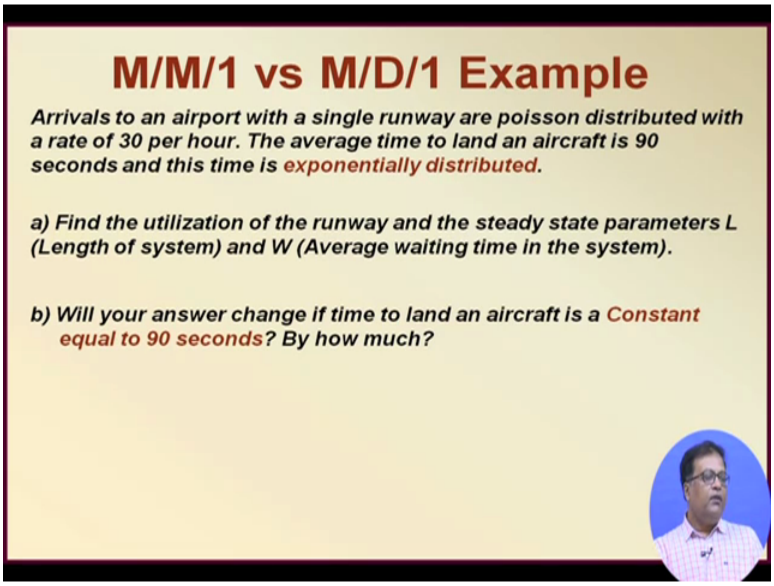
(Refer Slide Time: 2:43)



So what really happens in M/D/1 model that is the since the service is deterministic, we find that the L q the expected number in the queue L q becomes half of the corresponding L q value of the M/M/1 system right, so for an M/D/1 system the L q suppose we call this L q as you know L q 1 or L q 2 then this one we call L q 1. So L q 2 equals to half of L q 1, L q 2 for M/D/1, L q 1 for M/M/1, so what will be the value of L q 2? L q 2 will be half of Rho square by 1 – Rho right, so it should be Rho square by 2 into 1 – Rho. So what will be the L 2, that is expected number in the system expected number in the system, we know this is equal to L q 2 + Rho.

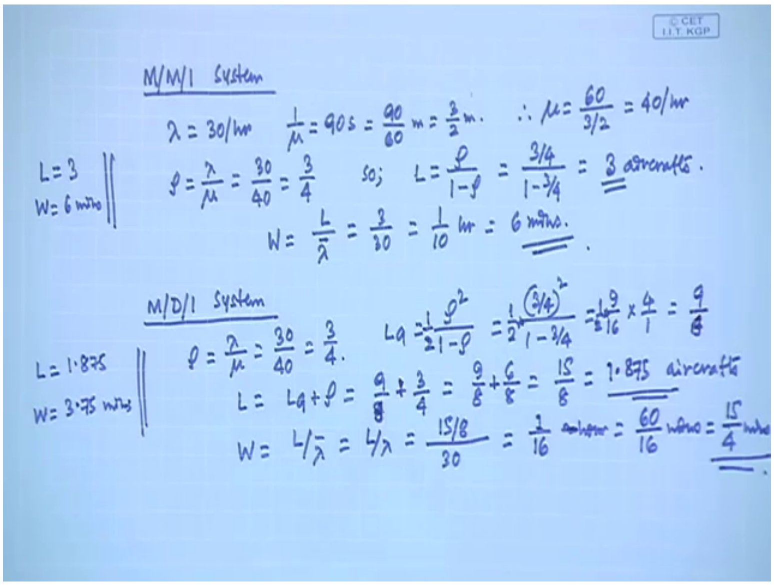
So it should be Rho + half into Rho square by 1 – Rho right, so that will be the L 2 and other parameters that is W 2 equals to L 2 by Lambda bar and W q 2 is L q to by Lambda bar right so those are the waiting times. So you can find out all the 4 parameters basic parameters of a queuing system that is expected number in the in the system, expected number in the queue, waiting time in the system and waiting time in the queue, once we know the L q value right, so L q will be exactly half of that of M/M/1 system. So this much result if you remember then you can easily work out all this right. So again once we will look here, L q is half and L equals to L q by Rho and this is how they work right, so this is M/D/1 model.

(Refer Slide Time: 5:59)



Now let us take an example, arrival to an airport with a single runway are Poisson distributed with a rate of 30 per hour. The average time to land an aircraft is 90 seconds and this time is exponentially distributed right. So find the utilisation of the runway and the steady-state parameters L and W average waiting time in the system. Will your answer change if time to land an aircraft is a constant equal to 90 seconds? And if so by how much? Is it all right?

(Refer Slide Time: 6:53)



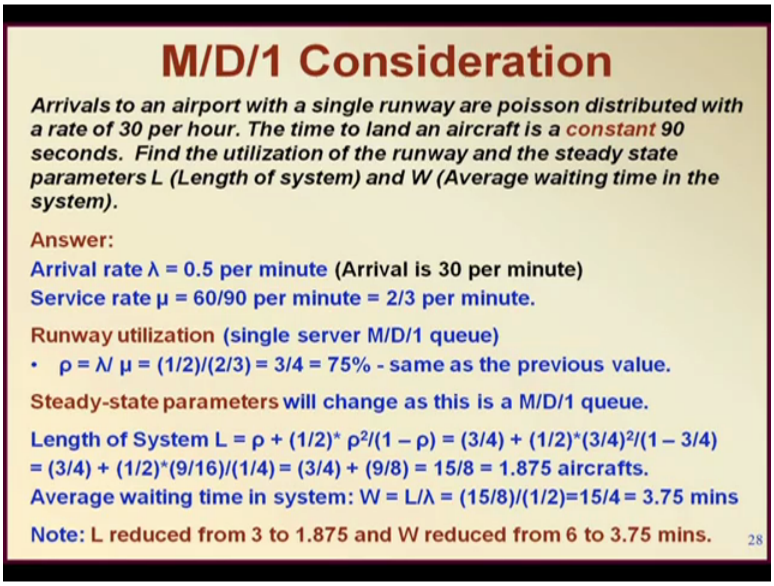
So let us look at this particular problem, so in this case the first is let us write down the M/M/1 system , we have got Lambda equals to 30 per hour and 1 by Mu equals to 90 seconds right 90 seconds equals to 90 by 60 minutes equals to 3 by 2 minutes so Mu equals to 60 by 30 by 2 equals to 40 per hour right, so that is the Lambda and Mu, so what will be the value of Rho? Rho will be Lambda by Mu that is the utilisation factor is 30 by 40 or 3 by 4 so that is the value of Rho. So Rho is obtained, now what will be the steady-state parameters L, length of the system L will be so L that is the total number in the both waiting and also in service, L will be Rho by 1 – Rho equals to 3 by 4 by 1 – 3 by 4 equals to 3 because 1 by 4 other side.

And what will be the value of W, W will be L by Lambda bar but Lambda bar is nothing but Lambda because all Lambdas are equal so it should be 3 by 30 or 1 by 10 hours which is equal to 6 minutes right. So these are the results that we got, L equal to 3 and W equal to 6 minutes, so let us look at these are the results quick revision, arrival rate is 0.5, service rate is 2 by 3 then this Rho equals to 75 percent, L equal to 3 aircraft and average waiting time 6 minutes. What happens in the M/D/1? In M/D/1 system, Lambda and Mu remains the same so Rho is Lambda by Mu equals to 30 by 40 equals to 3 by 4. What will be the value of L q? L q equals to Rho square by 1 – Rho equals to 3 by 4 whole square + 1 – 3 by 4 equals to 9 by 16 into 4 by 1 equals to 9 by 4, so that will be the value of L q.

So what will be the value of L in this case? L will be L q + Rho equals to 9 by 4 + no it will be one half of so we have to take that so one half of so it will be half into then half into so it will become 9 by 8 right. So L q + Rho is 9 by 8 + 3 by 4 equals to 9 by 8 + 6 by 8 equals to 15 by 8 right, so 15 by 8 it will become 1.70 so 1.875 aircraft right so this is comes L. And what will be the W? W will be L by Lambda bar equals to L by Lambda equals to 15 by 8 divided by 30 right, so it will be 1 by 161 by 16 hours which is 60 by 16 minutes right, or in other words equal to 15 by 4 minutes is 3 point something.

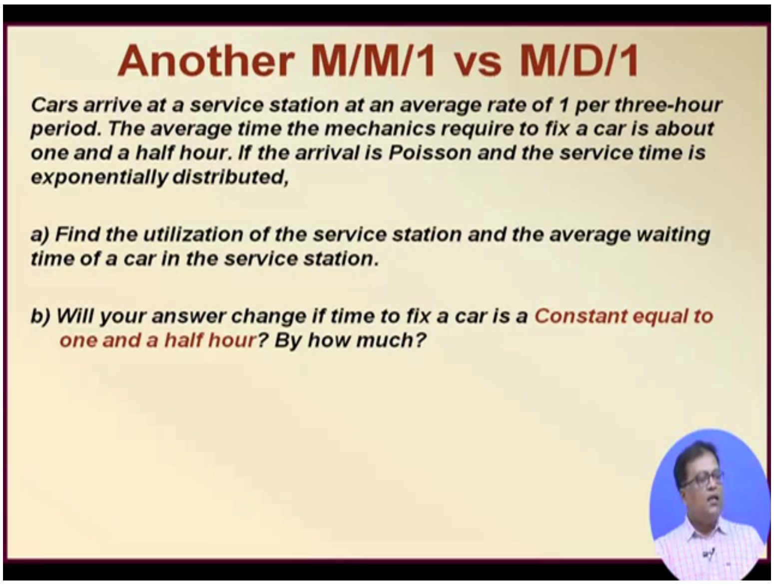
So look here the differences, the differences are very interesting to note whereas, for M/M/1 system L equal to 3 and W equal to 6 minutes, for M/D/1 system we get L equal to 1.875 and your W equals to 3.75 minutes. So what is the advantage that we got? If we can make the service deterministic in other words, if we can remove the variation from the service, we can actually bring down the values of the what you call the number in the system as well as waiting time substantially right. So if you really want to improve the queuing parameters, one method is really add more servers (())(13:36) but if you can also remove the chance variations right, particularly if you can through computerisation or any other means, if you can really make the service more what is called deterministic, things will improve drastically right.

(Refer Slide Time: 14:13)



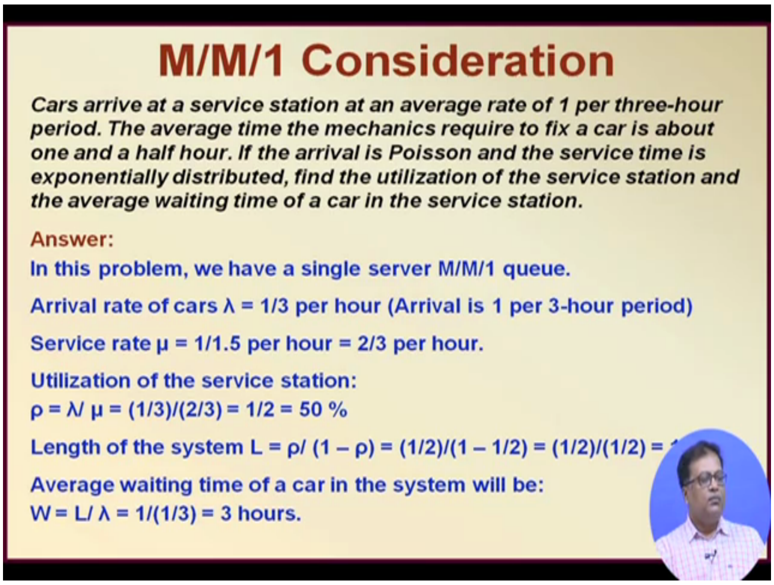
So this is what is shown in the M/D/1 consideration, arrival to an airport with a single runway Poisson distribution and time to land is constant time t is a constant, so here Rho is 75 percent, so look here how the L is calculated. Rho + half into Rho square by 1 – Rho, so Rho square by 1 – Rho is usually L q, take half of it and then add Rho to get the L because it is an M/D/1 system so it comes to 1.875 and average waiting time is 3.75 so this is the reduction or advantage that we have got from M/D/1 consideration.

(Refer Slide Time: 14:50)



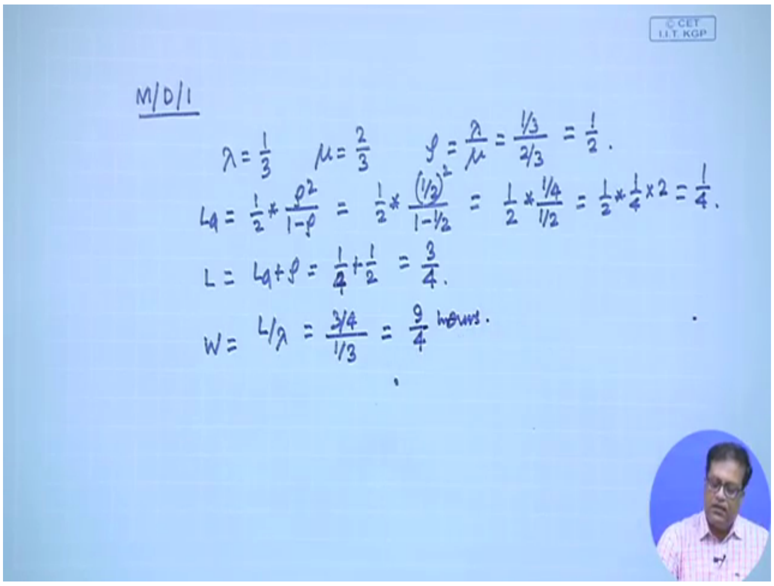
Just for example let us take one more example, the cars arrive at a service station at an average rate of one per 3 hour period right. The average time the mechanics required to fix a car is about 1 and a half hour. If the arrival is Poisson and the service time is exponentially distributed, find the utilization of the service station and average waiting time of a car in the service station. Will your answer change if time to fix a car becomes constant equal to 1 and a half hour and if so by how much? Right. So in this case what is happening, the utilisation of the service time and the average waiting time of the car in the service station, so let us look at these figures.

(Refer Slide Time: 15:41)



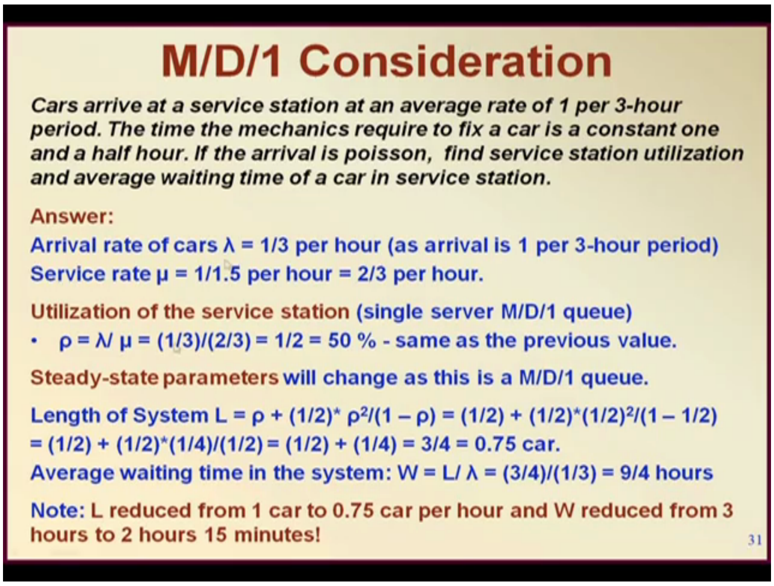
So here what happens in this problem, it is a single server M/M/1 queue, the arrival rate of cars the Lambda is 1 per 3 hours so 1 by 3 per hour right and the service rate is 2 by 3 per hour, so what is the utilisation? Utilisation Rho is equals to Lambda by Mu is equal to half 50 % right, and what will be the length of the system? The length of the system will be L which is Rho by 1 – Rho half by 1 – half, it comes out to be 1 car. And average waiting time of the car in the system W equals to L by Lambda equals to 3 hours, what happens if we do what is known as the M/D/1 consideration, right?

(Refer Slide Time: 16:56)



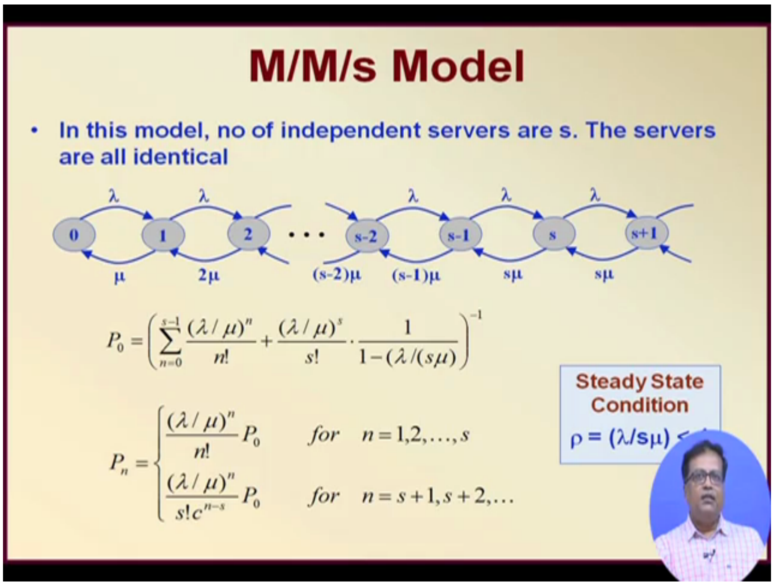
We have seen that in this M/D/1 consideration for this problem we have seen the Lambda equals to 1 by 3 and Mu equals to 2 by 3 so Rho equals to Lambda by Mu equals to 1 by 3 by 2 by 3 equals to half so that is equal to Rho. So what will be the L q in this case? The L q will become half of Rho square by 1 – Rho equals to half into half square by 1 – half equals to half into 1 by 4 by 1 by 2 equals to half into 1 by 4 into 2 is it not? so it is equal to 1 by 4, so this is L q so what will be L? L equal to L q + Rho equals to One fourth + half equals to 3 by 4. Compare this L, the earlier L was one car, now L becomes 3 by 4 cars right, and what is the waiting time? Waiting time will be L by Lambda equals to 3 by 4 divided by 1 by 3 equals to 9 by 4 Right, so it will be 9 by 4 hours, so let us look at all these results once again how we get that once again

(Refer Slide Time: 18:36)



That Rho equals to Lambda by Mu 50 percent, length of the system L comes out to be Rho into half Rho square by 1 – Rho 0.75 car right, 3 by 4 cars we have got. And average waiting time in the system W is L by Lambda is 3 by 4 divided by 1 by 3 equals to 9 by 4 hours. So what really happens that L reduced from one car to 0.75 car per hour and W reduced from 3 hours to 2 hours 15 minutes right. So there is a substantial you know improvement that we can find out if we can really can stabilise the service or reduce the variation particularly in the service time. Let us now move over to second model that is the M/M/S model, what is really an M/M/S model?

(Refer Slide Time: 19:34)



M/M/S model is an multiple server model, so in this multiple server model obviously certain things will be different, what will be different? If you look at the red diagram look at the red diagram carefully, the first when there is 0 you know when there is the state 0 then there could be an arrival and arrival will be always Lambda, all these things will be Lambda there is no issue. But what would be the service? When there is only one person in the system, the service could be Mu but when there are more than one person in the system the service could be 2 Mu right. And obviously when the servers are s – 2, it will be s – 2 Mu, when s – 1, it will be s – 1 Mu and when there is you know s server then onwards it will be s Mu.

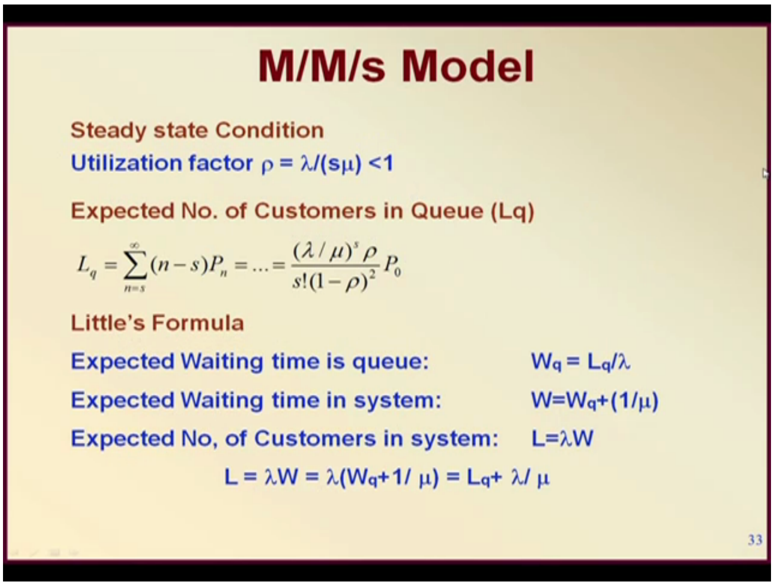
So since really the system behaves differently when there are s number of servers and when there are less than s number of people in the system. When there are s number of servers in the system and there are more than s number of people in the system then the service rate will be a constant s Mu, there is no issue there. But when the number of people in the system will be less than S, at that time obviously we cannot serve s number of people, we shall be able to serve only less than s number of people exactly to the number that is present. So therefore the analyses also will be therefore divided in 2 parts, one is what happens if there are more than s number of people in the system and what happens if there are less than s number of people in the system.

So if you look at this way, therefore the formula really consists of 2 parts; one part is the you know I am really not going into the derivation because it will be a elaborate detail thing all that you have to do, again but the process is the same. Process is the same that you have to do the equilibrium that means at 0 the Lambda into P 0 equals to Mu into P 1. But at one what happens, what goes out is Lambda into P 1, Mu into P 1, what comes in, Lambda into P 0 2 Mu not really, earlier we had taken only Mu, 2 Mu into P 2 right that is the balance at 1. Similarly if you keep doing the balance at different points and then you know to simplify then we can get the all these results. But really speaking, really do not have to remember all these results, we simply have to remember a few results, what are those few results?

The first most important result is that of P 0, the P 0 comes out to be n equals to 0 to s – 1 Lambda by Mu to the power n by factorial n that is the first term right. And the second term is Lambda by Mu to the power s by factorial s 1 by 1 – Lambda by s Mu. Look here the Lambda by s Mu is actually nothing but the Rho which is our traffic intensity, how it compares with the M/M/1 system mission, in an M/M/1 system the traffic intensity was Lambda by Mu and in M/M/s system the Lambda becomes Lambda by s Mu right. So the basic difference in M/M/1 system the traffic intensity was Rho equals to Lambda by Mu, here we get Rho equals to Lambda by s Mu so that is one of the difference.

The second difference the P 0 there was a very simple calculation, it was 1 – Rho but here it becomes an involved calculation, you have to sum over 0 to s – 1 Lambda by Mu to the power n by factorial n and Lambda by Mu to the power s by factorial s, 1 by 1 – Rho right, in this case however, Rho equals to Lambda by s Mu so we differentiate by Rho and Lambda by Mu in M/M/s system. What will be the P n, P n will be Lambda Mu to the power n by factorial n into P 0 till n reaches s right. And beyond this, this factor is Lambda by Mu to the power n factorial s into C to the power n – s right, I mean sorry this C is nothing but S, s to the power n – s P 0 right, so these are the formula that we shall use for this computing the probabilities.

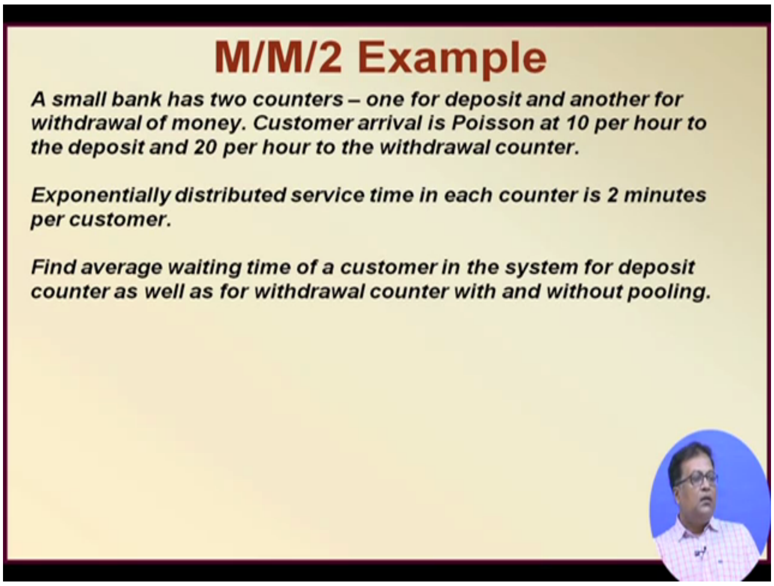
(Refer Slide Time: 24:48)



Afterwards, please remember there is only one formula that we need to remember if you want the other parameters, what are those other parameters? The first only formula that we need to remember is that of L q. The L q in this case is n equals to s to infinity n – s P n and it comes out to be Lambda by Mu to the power s Rho by factorial s 1 – Rho whole square into P 0. Before we proceed we must also remember that this P 0 is all these terms to the power – 1. Sometimes we forget that to the power – 1, you know the sum over the first-term, sum over the second set of things and addition of this which is sum over the all the remaining ones to the power – 1 and this – 1 must be clearly remembered right so that is the value for P 0.

And then as I said in M/M/s model the L q formula is what we really require and rest of the things can be obtained by little's formula but do remember that some of the simple results like L q equal to you know those L + Rho et cetera we should be very carefully you know should be using those results right, because here Rho is not Lambda by S Mu, here Rho will be Lambda by Mu because you know the other part of the formula is more robust that is W equals to W q + 1 by Mu right. So from there L equals to Lambda W there you will get that L and L q relation right. So do remember that Rho equals to Lambda by s Mu, L q equals to Lambda by Mu to the power s Rho divided by factorial S into 1 – Rho whole square P 0 right, so these are the sum M/M/s formula we shall use. Now let us look at a particular example of an M/M/2 system.

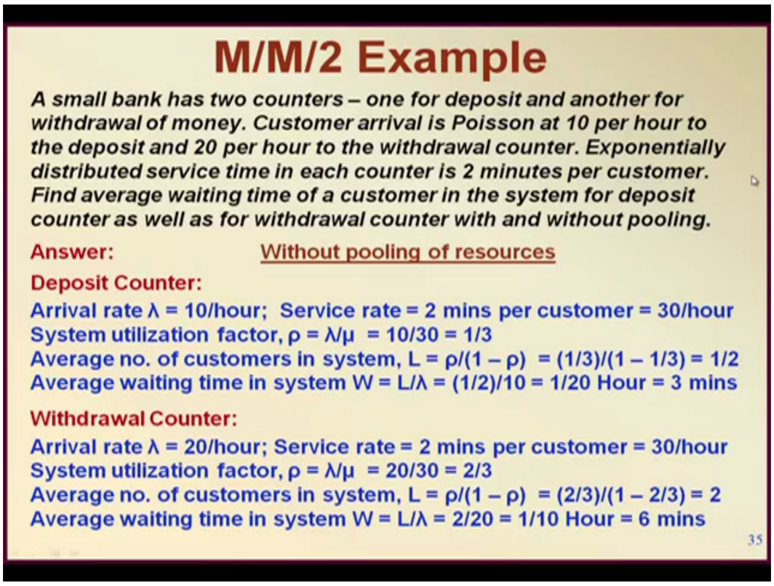
(Refer Slide Time: 26:58)



You see, a small bank has 2 counters; one for deposit and another for withdrawal of money. Customer arrival is Poisson at 10 per hour to the deposit and 20 per hour for the withdrawal counter. Exponentially distributed service time in each counter is 2 minutes per customer. Find average waiting time of a customer in the system for deposit counter as well as for withdrawal counter with and without pooling. Now you see, this cooling becomes an important parameter essentially what it means that if we pool the resources or in other words if we make the resources together, that means suppose both the counters can do both. Usually what happens the withdrawal counter can only do withdrawal of money and the deposit counter can only do deposit of money.

But if both can do both that means what? You can go any of the counters, in the first counter can do both deposit and withdrawal and in the withdrawal counter you can do again both deposit and withdrawal, what will be the advantage? Sometimes you find that deposit counter is busy but withdrawal counter is free and the reverse at certain point of time, right. You can utilise those ideal periods and therefore you can perhaps achieve better results, let us see another example.

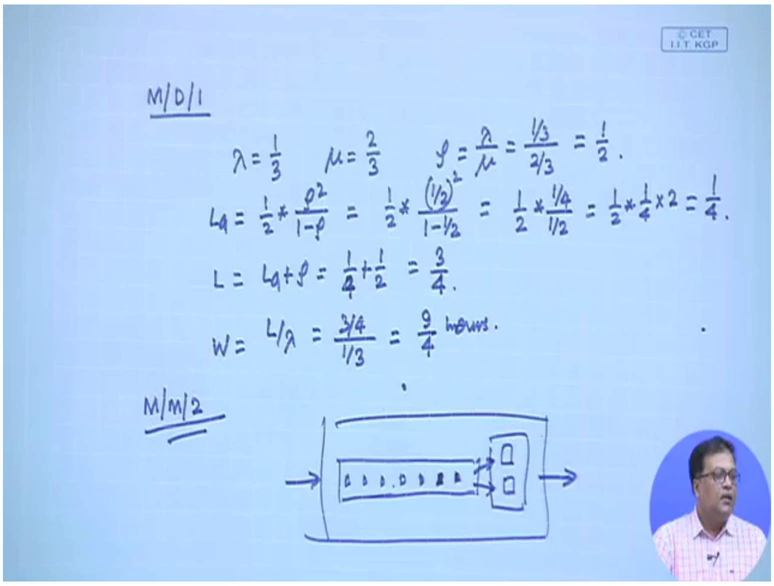
(Refer Slide Time: 28:38)



What happens without pooling of resources? It is the usual calculation, arrival is 10 per hour, service rate is 2 minutes that is 30 per hour, utilisation factor Rho is Lambda by Mu that is 1 by 3. What will be that average number of customers in the system L equals to Rho by 1 – Rho, it comes out to be half and average waiting time 3 minutes. Similar calculation for withdrawal counter becomes average number becomes 2 and average waiting time becomes 6 minutes. So summarise, if you do not pool you can basically it means two M/M/1, two separate M/M/1 systems right, two separate M/M/1 system, in the first in the deposit counter the average number of customers is half, average waiting time is 3 minute and in the withdrawal counter average number of customers is 2 and average waiting time is 6 minutes.

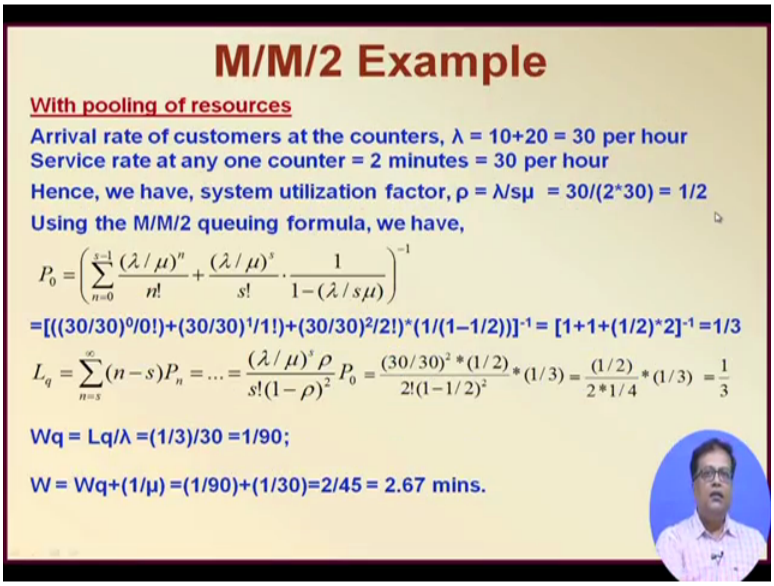
Now when you pool the resources then what happens? Now you see the arrival rate of customers at the counter now becomes 30 minutes. Please remember one thing that is let me draw it here that is in an M/M/1 system, let us call at M/M/2 system, let us say an M/M/2 system right.

(Refer Slide Time: 30:03)



In an MM 2 system this is the server and there are 2 servers right, this is the queuing system but then we still assume that there is a single queue right, all these people are waiting, they are waiting in a single queue this is how we, we are not really if there are 2 separate lines in front of the 2 counters then that system is nothing that is not a you know M/M/s system, that is essentially nothing but two M/M/1 system, please remember that obviously. But suppose see these people if they are actually maintaining to queues but whenever one is free, anybody can go from anywhere then virtually it becomes a single queue right, so must remember that. The queue is one, there are 2 servers, people are coming here, they are going out there and from this single queue they are joining either this server or that server depending on what is free so that is the essential idea.

(Refer Slide Time: 31:13)

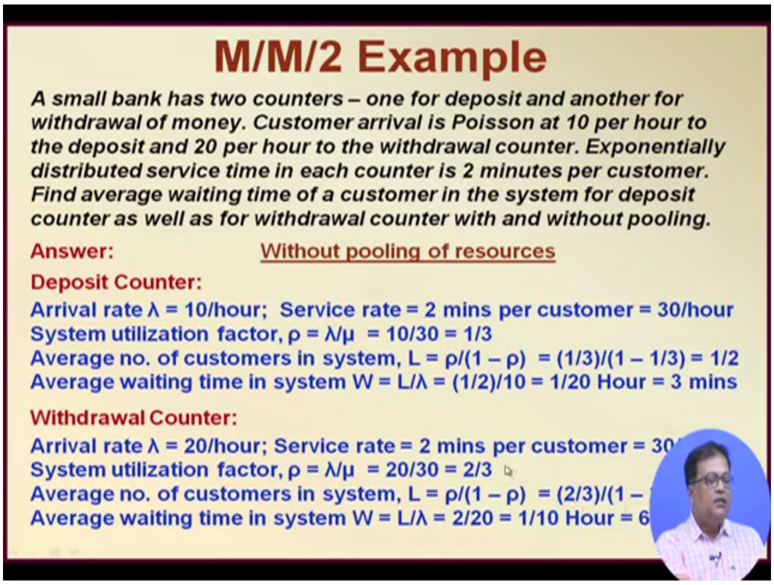


Now let us look at the calculations, so here Lambda becomes the 10 + 20 equals to 30 that is 30 and service rate at anyone counter becomes 2 minutes that is 30 per hour so that is what are the given values. So what will be the utilisation? Utilisation will be Lambda by s Mu equal to half. Now let us use the formula, the formula is P 0 is n equal to 0 to s – 1 Lambda by Mu to the power s by factorial s 1 by 1 – Lambda by S Mu. So you see the first term is then when you put 0 that will become 30 by 30 to the power 0 by 0 factorial, so this term is nothing but 1 because you see something to the power 0 is 1 and factorial 0 is 1, so first term will be one all the time. The second term is 30 by 30 to the power 1 divided by 1 because here Lambda by Mu are both are same so that is also coming out to be 1.

The third term is 30 by 30 square you know if you see this is the third term 30 by 30 this is the 30 by 30 to the power 2 by factorial 2 and 1 – 1 – half because this is the half is our utilisation factor. So for a small number of server this formula is not that difficult, it is not really that difficult, you see first term will be one, second term will be really Lambda by Mu to the power s by factorial s and the third term will be the Lambda by Mu to the power 2 because third term will be that is in this case only 2 servers by factorial 2 and multiply by 1 – 1 by 1 – the traffic intensity, now system utilisation factor. So it will be 1 + 1 + half into 2, which is again 1, 1 + 1 + 1 equals to 3 to the power – 1 that is 1 by 3 so that is comes out to be P 0.

And what is L q? L q will be Lambda by Mu to the power s, 30 by 30 square into Rho which is half and then factorial 2 by 1 – half whole square into 1 by 3, which comes out to be 1 by 3 so both P 0 and L q comes out to be 1 by 3. And W q therefore will be L q by Lambda is 1 by 90 and W becomes what is known as 2.67 minutes.

(Refer Slide Time: 34:01)



So look here, the earlier hour W was 3 minutes and 6 minutes, after polling the W becomes 2.65 minutes right. So you see that is the kind of advantage that one can really get, so earlier people were waiting 3 minutes in one queue and 6 minutes in another queue and later on people are waiting in either queue maximum 2.67 minutes, so this is what the advantage we get out of pooling of resources. We shall discuss more on pooling of resources and things of M/M/2 system in our next class, thank you very much.