Course on Decision Modeling
Professor Biswajit Mahanty
Department of Industrial and Systems Engineering
Indian Institute of Technology Kharagpur
Module 04
Lecture No. 18
M/M/s and M/M/Infinity Models

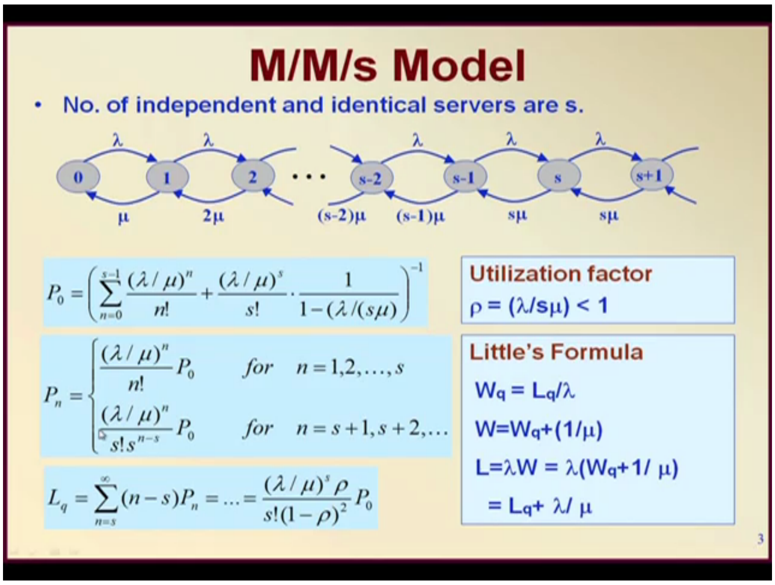
Today let us continue from where we discussing the queuing theory concepts, so far we have discussed the basic necessity that why queuing networks are required. And thereafter we have seen the birth and death process, the different types of giving examples related to the distribution likes exponential and Poisons distribution and thereafter we also solved problems related to let us say mostly the MM 1 type of queues, but one problems we have also solved from multiple servers.

(Refer Slide Time: 01:04)

Module 18

M/M/s and M/M/∞ Models

2



M/M/s Model

- No. of independent and identical servers are s.

λ
λ
λ
λ
λ
0
1
2
• • •
s-2
s-1
s
s+1
μ
2μ
(s-2)μ
(s-1)μ
sμ
sμ

$$P_0 = \left(\sum_{n=0}^{s-1}\frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!}\cdot\frac{1}{1-(\lambda/(s\mu))}\right)^{-1}$$

$$P_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!}P_0 & for \quad n = 1,2,\ldots,s \\ \frac{(\lambda/\mu)^n}{s!s^{n-s}}P_0 & for \quad n = s+1, s+2,\ldots \end{cases}$$

$$L_q = \sum_{n=s}^{\infty}(n-s)P_n = \ldots = \frac{(\lambda/\mu)^s \rho}{s!(1-\rho)^2}P_0$$

Utilization factor

$\rho = (\lambda/s\mu) < 1$

Little's Formula

$W_q = L_q/\lambda$

$W = W_q + (1/\mu)$

$L = \lambda W = \lambda(W_q + 1/\mu)$

$= L_q + \lambda/\mu$

3

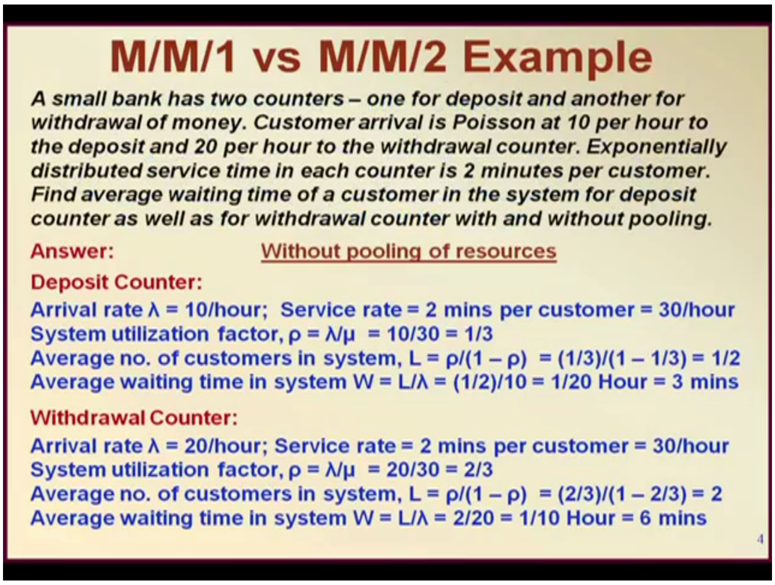
Now today let us look further into 2 categories of queuing system, particularly multiple servers MMS and a special case of the multiple servers which can be called the infinite server or the self-service model. Now already these equations we have seen in our previous class that when there are multiple servers then the formula are little involved and however not difficult to compute only thing the formula is little bit we need to remember.

But other than that this formula see but again you know there are only 2 things which one need to recall, one is the value for P 0 that formula is given here that is the very first one, and the formula for LQ, right. So must the person knows the probability of 0 persons in the system that is P 0 and also the formula for LQ that is the expected number in the queue, using this 2 formula one can find all the other formula by using the Little’s formula. You know WQ LQ by lambda, W WQ plus 1 by Mu and L lambda W that is lambda WQ plus 1 by Mu.

But one thing must be remember here that the utilization factor in this case is not lambda by Mu but lambda by S Mu, so this factor has to be remembered very-very clearly. So some books you will find which you will say rho equal to lambda by Mu, is it alright? But in that case that rho is not utilisation factor, the way we have used rho here it is taken from the Helier Liberbens book and rho is used as lambda by S Mu which is utilization factor.

Those books which writes lambda by Mu equal to rho they just substituted lambda by Mu by that factor and they do not mean rho by utilisation factor. So it should be also remember therefore in particularly the LQ formula where it is written lambda by Mu to the power s rho by factorial s 1 minus rho whole square P 0, that is nothing but the utilisation factor that is lambda by s Mu so this fact has to be remembered very carefully.

(Refer Slide Time: 03:54)



M/M/1 vs M/M/2 Example

A small bank has two 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 to 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.

Answer: Without pooling of resources

Deposit Counter:

Arrival rate λ = 10/hour; Service rate = 2 mins per customer = 30/hour
System utilization factor, ρ = λ/μ = 10/30 = 1/3
Average no. of customers in system, L = ρ/(1 – ρ) = (1/3)/(1 – 1/3) = 1/2
Average waiting time in system W = L/λ = (1/2)/10 = 1/20 Hour = 3 mins

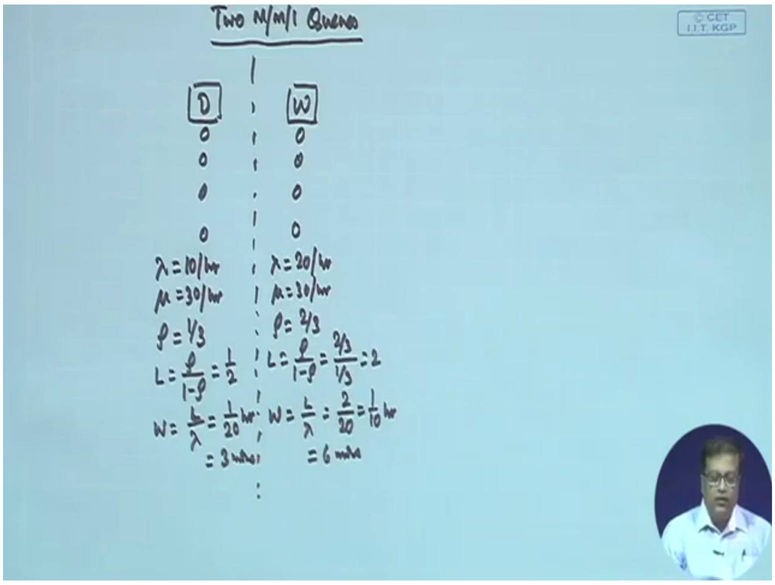
Withdrawal Counter:

Arrival rate λ = 20/hour; Service rate = 2 mins per customer = 30/hour
System utilization factor, ρ = λ/μ = 20/30 = 2/3
Average no. of customers in system, L = ρ/(1 – ρ) = (2/3)/(1 – 2/3) = 2
Average waiting time in system W = L/λ = 2/20 = 1/10 Hour = 6 mins

4

Now a problem that we had already solved about small bank that has 2 counters, one for deposit another for withdrawal and arrival in both is Poisson process at 10 per hour and while the service is you know 10 per hour for deposit and 20 per hour for withdrawal and exponentially distributed service time is 2 minutes per customer for each of the counters. So usually for such problems we like to know the average waiting time in the system for the both counters and we have already solved this problem for MM 1.

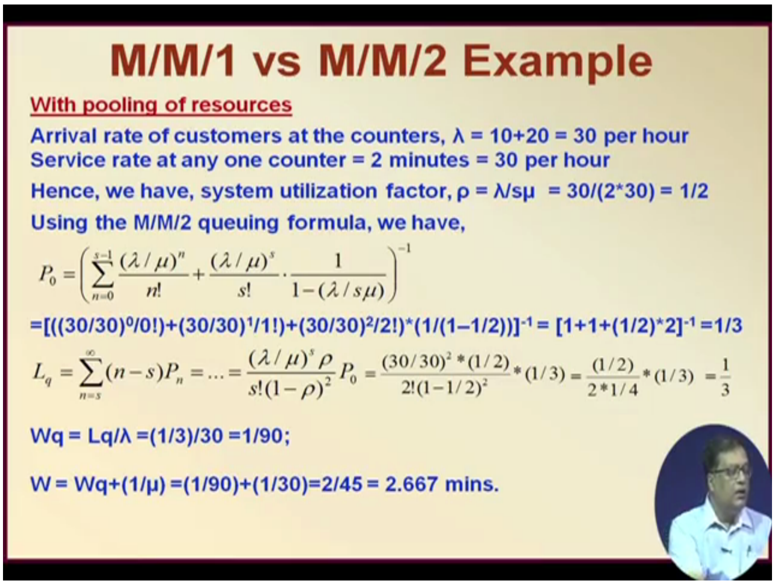
(Refer Slide Time: 04:39)



So you see this is the first case where we think that this is my deposit counter, there is a queue formed in front of the deposit counter, there is a withdrawal counter the queue formed in front of the withdrawal counter. So these 2 are separate processes and for these separate processes we already found that for this one lambda is 10 per hour and Mu is 30 per hour, so we had rho equal to 1 by 3, we had L equal to rho by 1 minus rho equal to half and we also have W equal to L by lambda which has come up to be, because lambda is 10, one by 20 hour that is equal to 3 minutes.

So this we had seen already for the deposit counter and for the other counter this is 20 per hour, Mu equal to 30 per hour, rho equal to 2 by 3, L equal to rho by 1 minus rho equal to 2 by 3 by 1 by 3 equal to 2 and W equal to L by lambda equal to 2 by 20 equal to 1 by 10 hour equal to 6 minutes. So this calculation if we had already done and this is available to us in our previous lecture but just repeated here, but one thing must remember that these are 2 MM 1 queues, right.

(Refer Slide Time: 07:03)



M/M/1 vs M/M/2 Example

With pooling of resources

Arrival rate of customers at the counters, λ = 10+20 = 30 per hour
Service rate at any one counter = 2 minutes = 30 per hour

Hence, we have, system utilization factor, ρ = λ/sμ = 30/(2*30) = 1/2

Using the M/M/2 queuing formula, we have,

$$P_0 = \left(\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \cdot \frac{1}{1-(\lambda/s\mu)} \right)^{-1}$$

$$=[((30/30)^0/0!)+(30/30)^1/1!)+(30/30)^2/2!)*(1/(1–1/2))]^{-1} = [1+1+(1/2)*2]^{-1} =1/3$$

$$L_q = \sum_{n=s}^{\infty}(n-s)P_n = \ldots = \frac{(\lambda/\mu)^s \rho}{s!(1-\rho)^2} P_0 = \frac{(30/30)^2 * (1/2)}{2!(1-1/2)^2} * (1/3) = \frac{(1/2)}{2*1/4} * (1/3) = \frac{1}{3}$$

Wq = Lq/λ =(1/3)/30 =1/90;

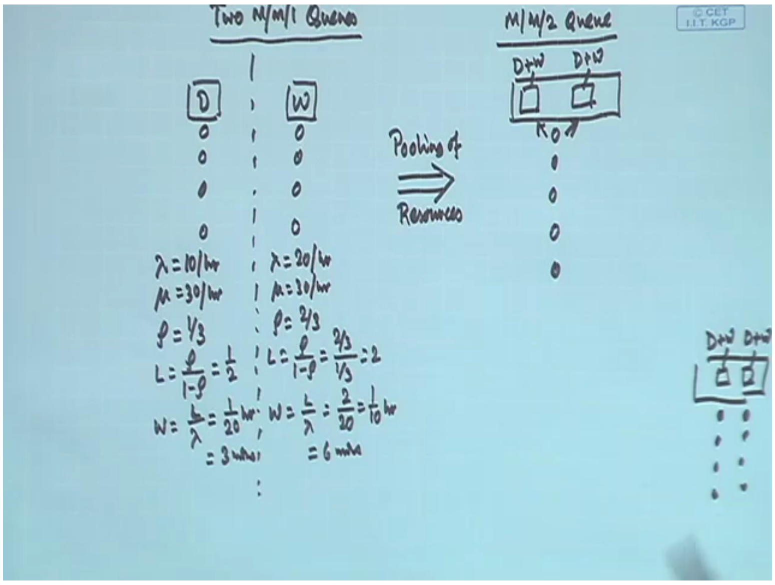
W = Wq+(1/μ) =(1/90)+(1/30)=2/45 = 2.667 mins.

Now when we use what is known as MM 2 queue, what exactly we do? You know we do combine that 2 and we have a system which is an MM 2 queue however we are getting this MM 2 queue from the 2 MM 1 queue is a process what is known as pooling of resources, right. What is the pooling of resources? Now there is a server system where it is and each can do deposit plus withdrawal, right, so these are 2 counters and both can do deposit as well as withdrawal.

Now question is that this is something very tricky question, how is are they are 2 different queues in front of the 2 counters or there is a single queue. You see as far as the literature goes we assume that there is a single queue, but what is the difference between these process and let us say another process just think about this, these are the 2 counters, there are 2 different queues in front of the 2 but both can do both this is also can do deposit plus withdrawal and both the counters can deposit plus withdrawal, what is the difference between this 2?

You see what happen in this particular case that you know suppose one of the queue is empty then will the person keep waiting here, I mean will the other people who are you know waiting in a long queue in the other counter they will simply come here and get their service. So in a way as if there are single queue only thing that queue discipline maybe different, because say things are different as far as the individual concerned. Here the person knows that so many peoples are before me, but here this person knows only one person is before me and I can get my service once the previous service is over.

(Refer Slide Time: 09:31)

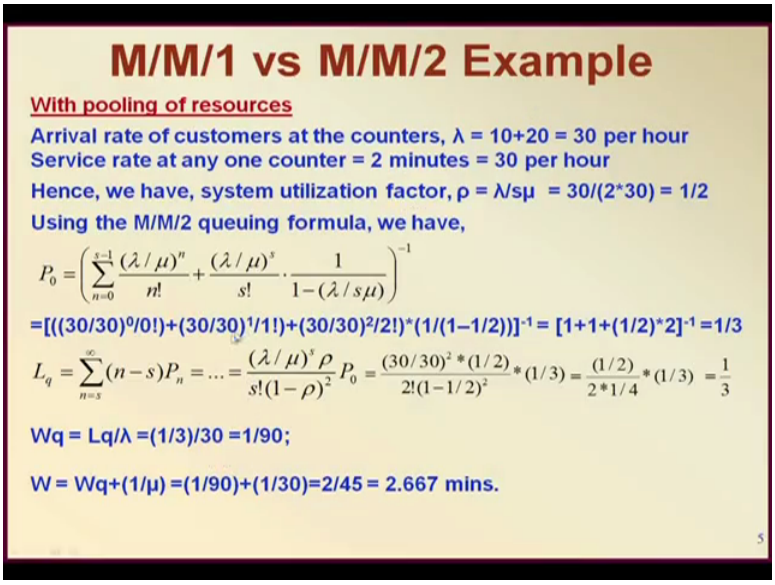


So there is a difference from customer point of view, but as you remember already I have told that the performance variable values steady-state performance variable values what are some performance variable values? The L, LQ, W, WQ, they are independent of queue discipline, what kind of queue discipline? The kind of queue discipline that does not affects the process.

Now this queue discipline like FCFS-SIRO service in the random order and last in first out, this kind of system they do not what is known as you know they are not changing the system input or service processes therefore those performance variable values are independent of queue discipline.

So under those situation that you know this L, LQ, W, WQ values they are not going to change, right. Depending on whether they are standing in 2 different queues or a single queue, only thing it make will make a difference in the service of individual person more about that later.

(Refer Slide Time: 11:08)



M/M/1 vs M/M/2 Example

With pooling of resources

Arrival rate of customers at the counters, λ = 10+20 = 30 per hour
Service rate at any one counter = 2 minutes = 30 per hour

Hence, we have, system utilization factor, ρ = λ/sμ = 30/(2*30) = 1/2

Using the M/M/2 queuing formula, we have,

$$P_0 = \left(\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \cdot \frac{1}{1-(\lambda/s\mu)} \right)^{-1}$$

$=[((30/30)^0/0!)+(30/30)^1/1!)+(30/30)^2/2!)*(1/(1-1/2))]^{-1} = [1+1+(1/2)*2]^{-1} = 1/3$

$$L_q = \sum_{n=s}^{\infty}(n-s)P_n = \ldots = \frac{(\lambda/\mu)^s \rho}{s!(1-\rho)^2} P_0 = \frac{(30/30)^2 * (1/2)}{2!(1-1/2)^2} * (1/3) = \frac{(1/2)}{2*1/4} * (1/3) = \frac{1}{3}$$

Wq = Lq/λ =(1/3)/30 =1/90;

W = Wq+(1/μ) =(1/90)+(1/30)=2/45 = 2.667 mins.

But let us see what happens with pooling of resources look at this slide, here this P 0 formula you know that will be given by you know this lambda by Mu to the power n factorial n, lambda by Mu factorial s and 1 by lambda by s Mu to the power minus 1 and this P 0 in our previous class we have seen that this P 0 comes out to be 1 by 3, I.

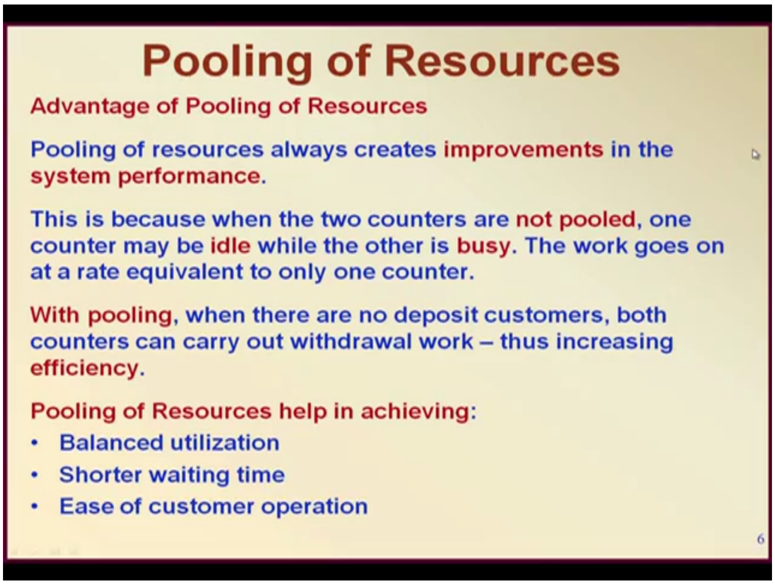
And what is the landfill here, the London will be equal to the 2 processes 10 plus 20 that is equal to 30 the Mu remains at 30 per hour, and what will be rho? Rho will be lambda by 2 Mu, why 2 Mu? Because there are 2 servers, so in this case sorry lambda by 2 Mu so this lambda is 30 and this will become 60, so rho will be half.

So in this case P 0 comes out to be 1 by the 3 that calculation I am not shown here, but it will become 1 by 3 and rho equal to half and computation of LQ using the formula that also comes out to be 1 by 3 hour, right sorry not hour 1 by 3 and W really comes out to be 1 by 3 into 60, that is because W is L by lambda, lambda is 30 so equal to 1 by 90 equal to how much it comes to be WQ is 1 by 90, so again we have to find that is WQ, W equal to WQ plus 1 by Mu equal to 2 by 45, right, and that comes to be 2.67 minutes.

Once again lambda is 30, Mu is also 30, P 0 is 1 by 3, rho is half, LQ comes out to be 1 by 3 from LQ we can find WQ by dividing LQ by lambda so it gets 1 by 90 and W then WQ plus 1 by Mu that is 1 by 90 plus 1 by 30 comes out to be 2 by 45 which is 2.67 minutes, right. So what is really happening that you see when the services were separate then one was taking 3 minutes and the other was taking 6 minutes, right, and when you combine them into a single MM 2 queue then we had you know the W comes out to be 2.67 minutes, that means average waiting time in the system reduces for both the counters.

And why it happens? It really happens because you know you see when a deposit counter is free, right, but withdrawal counter is very busy because there are lot of people waiting there is, you know this free counters are not helping you know people who are waiting in front of the withdrawal counter it does not help them, right. So but when what happens in the pooling? Since both the people can do deposit as well as withdrawal you know no counter will remain free as long as there are customers.

(Refer Slide Time: 15:25)



Pooling of Resources

Advantage of Pooling of Resources

Pooling of resources always creates improvements in the system performance.

This is because when the two counters are not pooled, one counter may be idle while the other is busy. The work goes on at a rate equivalent to only one counter.

With pooling, when there are no deposit customers, both counters can carry out withdrawal work – thus increasing efficiency.

Pooling of Resources help in achieving:

- Balanced utilization
- Shorter waiting time
- Ease of customer operation

6

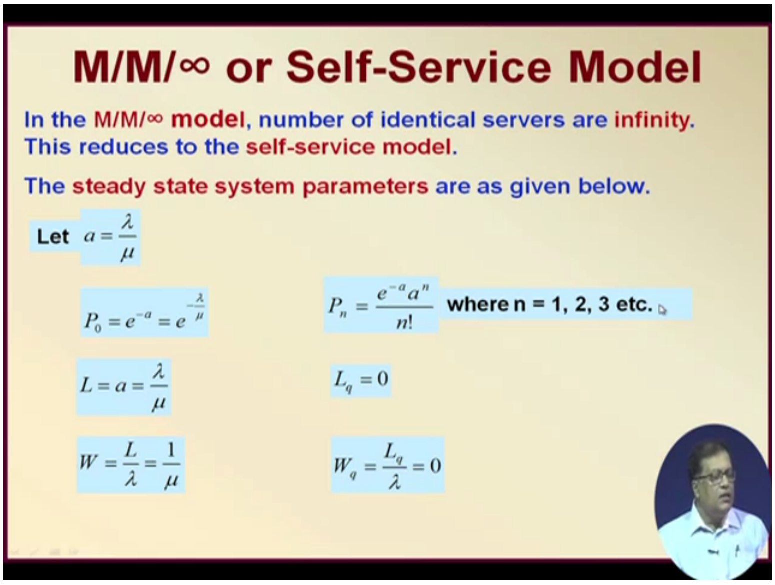
So this will reduce the free period of individual counters and pull resources in that sense and show better performance. So that is what is the advantage of pooling of resources you know that creates improvements in the system performance, this is because when the 2 counters not pooled one counter may be idle while the other is busy. The work was on at the rate equivalent to only one counter. With pooling when there are no deposit customers both counters can carry out withdrawal work that does increase efficiency.

So what pooling of resources really help in achieving? You know it helps in the first of all balanced utilisation. The utilisation of both the counters are possible and it does not happen that if see usually what will happen deposit counter only 10 per hour is the arrival, withdrawal counter it is 20 per hour, so therefore the deposit counter may remain free most of the time and whereas withdrawal counter will be busy all the time. But if both can do both then the load will be divided and both will be working and therefore the performance of both are going to improve, right, so that is the advantage that we get with pooling of resources.

The waiting time reduces the 2nd one shortened waiting time and finally also the ease of customer operation. If you remember in the early days when we used to buy railway tickets and you know there used to be counters which use to give only daily tickets, other counters will give something like madras tickets or Chennai tickets.

So you know you have to really wait in a particular counter for longtime, but you have to know what is the specific counter where you have to go and assume suppose you have to go to 5 directions and there are 5 different counters how difficult it is, but today you can join any queue, there are 5 counters join any of the 5 counters really and get your ticket. So that is the advantage of pooling of resources, balanced utilisation, shortened waiting time and ease of customer operation.

(Refer Slide Time: 17:40)



M/M/∞ or Self-Service Model

In the M/M/∞ model, number of identical servers are infinity. This reduces to the self-service model.

The steady state system parameters are as given below.

Let $a = \frac{\lambda}{\mu}$

$P_0 = e^{-a} = e^{-\frac{\lambda}{\mu}}$

$P_n = \frac{e^{-a}a^n}{n!}$ where n = 1, 2, 3 etc.

$L = a = \frac{\lambda}{\mu}$

$L_q = 0$

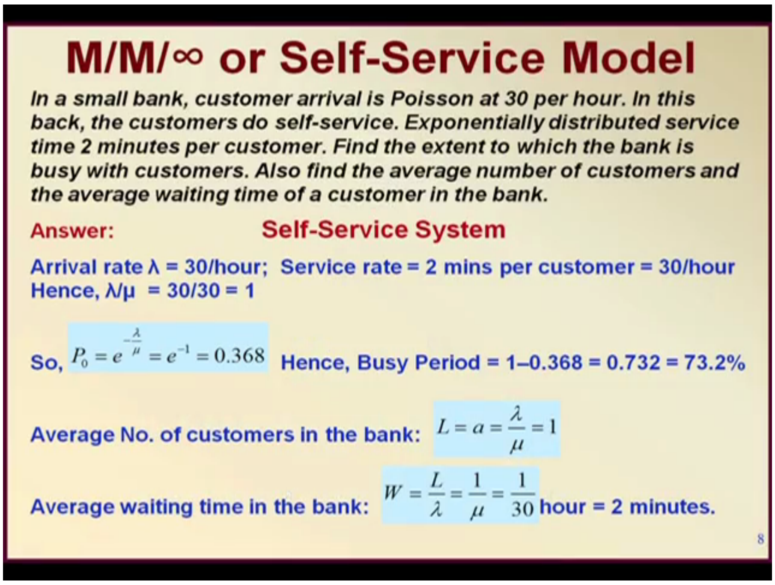
$W = \frac{L}{\lambda} = \frac{1}{\mu}$

$W_q = \frac{L_q}{\lambda} = 0$

A special case on this which is called the MM infinity or the self-service model, you see this is an extension of the multiple server model where an number of servers are infinite. Then what does it mean? It really means that nobody will going to wait, right, there is no question of wait. You come to the bank and there are infinite number of counters join any one of them, right. So it basically what? It is like self-service so when the self-service facility is available it is very clear that there is no expected number in the queue, so you can see LQ equal to 0, right, and also WQ will also be equal to 0.

So it becomes almost like a poison process and therefore we find that L will become lambda Mu, lambda is arrival rate, Mu is the service rate. I will not go into the derivations, but the formula for P 0 will become e to the power of minus lambda Mu, right. And other probabilities will be e to the power of minus a, a stands for Lambda Mu, a to the power n that is e to the power minus lambda by Mu, a to the power lambda by Mu by factorial n, so it is a poison process, right. And what will be the waiting time? Waiting time will be 1 by Mu. So that will be the difference in the formula, very simple formula really there LQ and WQ both are 0, W equal to 1 by Mu, L equal to lambda by Mu, that is what happens in a self-service process.

(Refer Slide Time: 19:35)



M/M/∞ or Self-Service Model

In a small bank, customer arrival is Poisson at 30 per hour. In this back, the customers do self-service. Exponentially distributed service time 2 minutes per customer. Find the extent to which the bank is busy with customers. Also find the average number of customers and the average waiting time of a customer in the bank.

Answer: Self-Service System

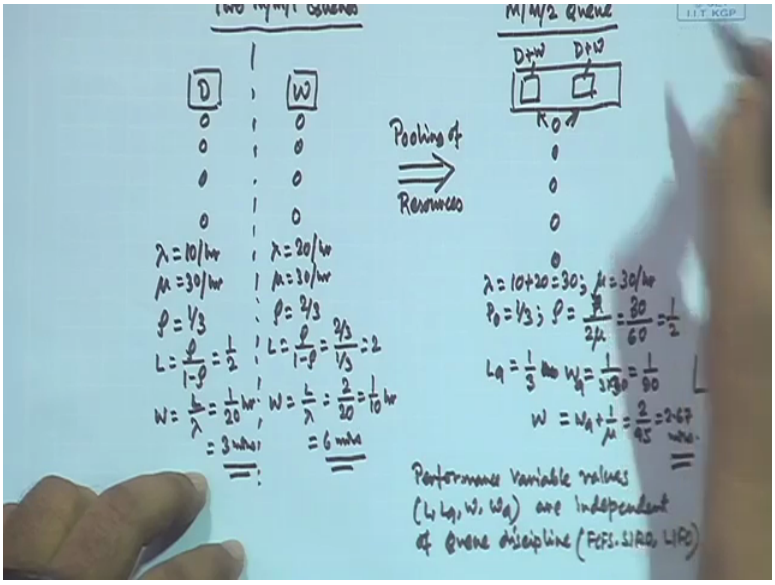
Arrival rate λ = 30/hour; Service rate = 2 mins per customer = 30/hour
Hence, λ/μ = 30/30 = 1

So, $P_0 = e^{-\frac{\lambda}{\mu}} = e^{-1} = 0.368$ Hence, Busy Period = 1–0.368 = 0.732 = 73.2%

Average No. of customers in the bank: $L = a = \frac{\lambda}{\mu} = 1$

Average waiting time in the bank: $W = \frac{L}{\lambda} = \frac{1}{\mu} = \frac{1}{30}$ hour = 2 minutes.

8

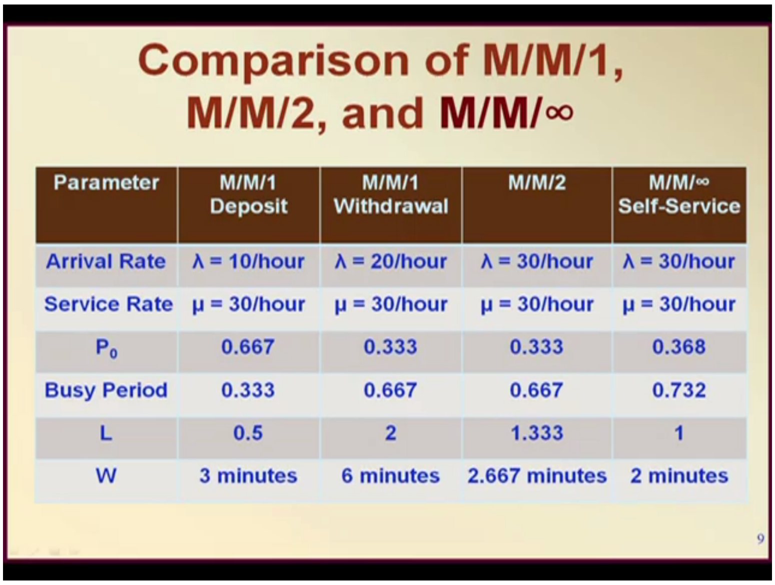


So for the same example you see what will happen in the self-service situation, the self-service situation in this case supposing this is our MM infinity queue let us write that also MM infinity queue, we have already seen what is known as the 2 MM 1 queue, MM 2 queue and suppose we have now MM infinity queue. In this case there are lambda equal to 30 per hour, Mu is also 30 per hour, right and this service system is like infinite queue any customers.

Say virtually there is a you know if I really draw a picture in this picture there are like a infinite servers, the customers can do you know virtually there is nobody waiting, everyone is as they come they get a separate place to really get their work done, so what will be the W in this case? The W formula already known is 1 by Mu, right and that is 1 by 30 that is equal to 2 minutes.

So look here how it changes when we had 2 different counters we had waiting time equal to 3 minutes and 6 minutes, when we pooled the resources and made it MM 2 right queue at that time we had the W becomes 2.67 minutes and really if we can make self-service, right, like something like you do it yourself on your computer I do not even come to bank and suppose it just takes 2 minutes to do one service then waiting time becomes simply 2 minutes, because that is the time you spend in getting the job done, it. Service time is how much? 2 minutes and that is the waiting time also because there is no queue, nobody waits in the queue so it is very simple, W becomes 2-minutes.

(Refer Slide Time: 21:51)



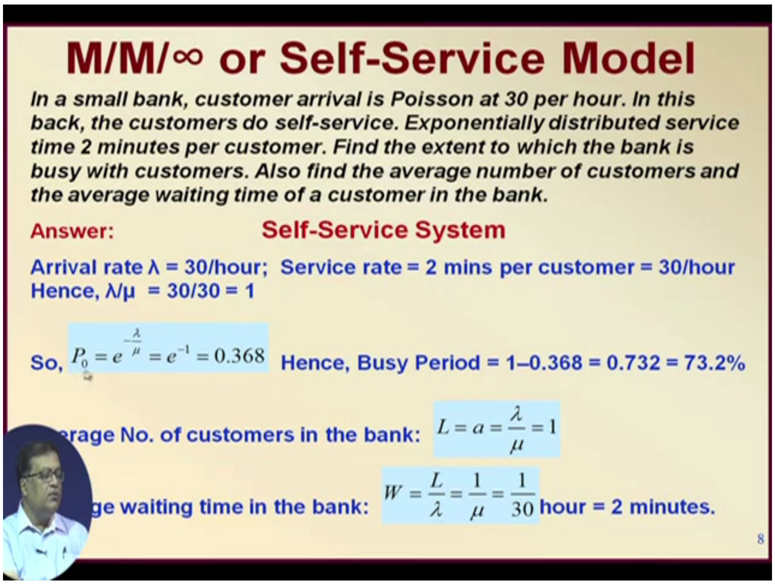
Comparison of M/M/1, M/M/2, and M/M/∞

Parameter	M/M/1 Deposit	M/M/1 Withdrawal	M/M/2	M/M/∞ Self-Service
Arrival Rate	λ = 10/hour	λ = 20/hour	λ = 30/hour	λ = 30/hour
Service Rate	μ = 30/hour	μ = 30/hour	μ = 30/hour	μ = 30/hour
P_0	0.667	0.333	0.333	0.368
Busy Period	0.333	0.667	0.667	0.732
L	0.5	2	1.333	1
W	3 minutes	6 minutes	2.667 minutes	2 minutes

9

So how do they compare? Here is a comparison chart just look what happens between MM 1 deposit, MM 1 withdrawal, MM 2 and finally MM infinity self-service, so let us look at all this. The arrival rate is 10 per hour, 30 per hour, 20 per hour and 30 per hour, 30-30, here also 30-30. Now P 0 formula this becomes 2 by 3 this becomes 1 by 3, because P 0 is 1 minus rho, rho is 1 by 3, so P 0 is 2 by 3, this is 0.333 and this is 0.368.

(Refer Slide Time: 22:32)



M/M/∞ or Self-Service Model

In a small bank, customer arrival is Poisson at 30 per hour. In this back, the customers do self-service. Exponentially distributed service time 2 minutes per customer. Find the extent to which the bank is busy with customers. Also find the average number of customers and the average waiting time of a customer in the bank.

Answer: Self-Service System

Arrival rate λ = 30/hour; Service rate = 2 mins per customer = 30/hour
Hence, λ/μ = 30/30 = 1

So, $P_0 = e^{-\frac{\lambda}{\mu}} = e^{-1} = 0.368$ Hence, Busy Period = 1–0.368 = 0.732 = 73.2%

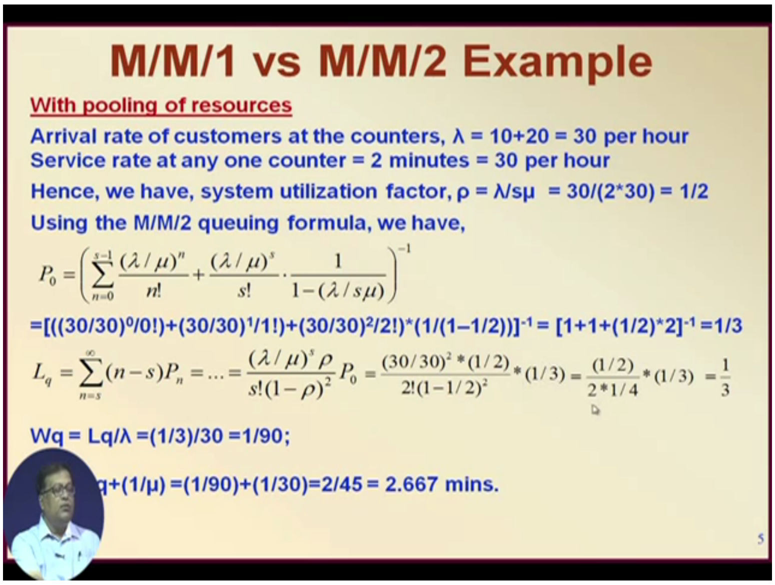
rage No. of customers in the bank: $L = a = \frac{\lambda}{\mu} = 1$

ge waiting time in the bank: $W = \frac{L}{\lambda} = \frac{1}{\mu} = \frac{1}{30}$ hour = 2 minutes.

8

How this 0.368 is obtained just look, P 0 e to the power minus lambda by Mu, e to the power minus 1 that is 0.368, Lambda by Mu is equal to 1, right. So you can see that this formula for P 0 really you know shows that the system is busy period is maximum, right, so this is something wrong this should be 0.632, 63.2 percent that is the busy period here this is something wrong. And whereas the other cases it is not busy that much but since how this thing is compensated really is by you know this infinite number of servers, because there are more number of servers, there is no waiting time that is how it is compensated.

(Refer Slide Time: 23:48)



M/M/1 vs M/M/2 Example

With pooling of resources

Arrival rate of customers at the counters, λ = 10+20 = 30 per hour
Service rate at any one counter = 2 minutes = 30 per hour

Hence, we have, system utilization factor, ρ = λ/sμ = 30/(2*30) = 1/2

Using the M/M/2 queuing formula, we have,

$$P_0 = \left(\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \cdot \frac{1}{1-(\lambda/s\mu)} \right)^{-1}$$

$=[((30/30)^0/0!)+(30/30)^1/1!)+(30/30)^2/2!)*(1/(1–1/2))]^{-1} = [1+1+(1/2)*2]^{-1} = 1/3$

$$L_q = \sum_{n=s}^{\infty} (n-s)P_n = \ldots = \frac{(\lambda/\mu)^s \rho}{s!(1-\rho)^2} P_0 = \frac{(30/30)^2 * (1/2)}{2!(1-1/2)^2} * (1/3) = \frac{(1/2)}{2*1/4} * (1/3) = \frac{1}{3}$$

Wq = Lq/λ =(1/3)/30 =1/90;

q+(1/μ) =(1/90)+(1/30)=2/45 = 2.667 mins.

5

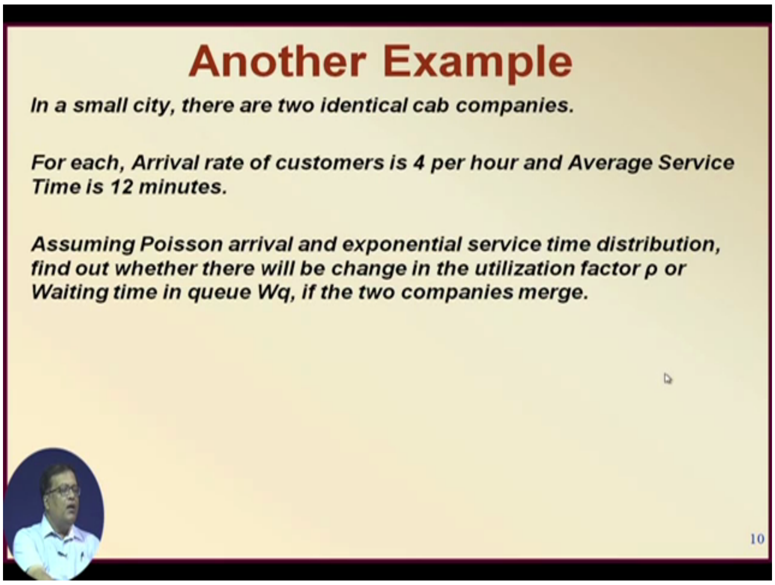
But one more interesting thing let us look, look at this MM 2 queue. The MM 2 queue the P 0 is 0.333, but busy period is 0.667, is it not. So but what was the latest go by quickly see what was the value of P 0, P 0 is 1 by 3 and the value of lambda is half that is sorry value of rho, system utilisation factor rho is lambda by s Mu that is half.

So system utilisation is 50 percent, right, so 50 percent time you know this system is not utilized fully stop but what is the busy period for MM 2? Is 66.7 percent. So how does it account for the remaining 16.7 percent? What is the really how this will be breached? This gap because severe rho lambda by s Mu which is half, because lambda is 30, Mu is 30, so 30 by 60, s is 2, 30 by 60 that you know the utilisation is only 50 percent.

But how the system is busy in 66.7 percent? Where this remaining 16.7 percent came? We have to understand that this gap is really coming from the consideration that you know while one counter is busy other could be ideal, right. So when it comes to utilisation then this utilisation is becoming half, right at that time, so the case could be both are busy, both are free, one is busy, other one is busy.

So when only one of the server out of the 2 are busy for those periods right the utilisation is really not full. So that is the difference, so that is why the busy period could be 66.7 percent that means that at least explains one counter is busy, however the you know utilisation is only 50 percent, so that point must be remembered.

(Refer Slide Time: 25:54)



Another Example

In a small city, there are two identical cab companies.

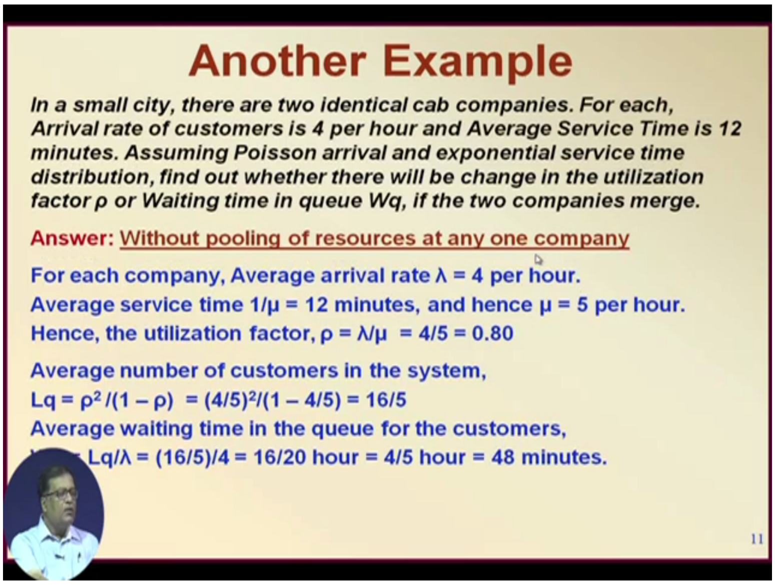
For each, Arrival rate of customers is 4 per hour and Average Service Time is 12 minutes.

Assuming Poisson arrival and exponential service time distribution, find out whether there will be change in the utilization factor ρ or Waiting time in queue Wq, if the two companies merge.

10

Look at another example, in a small city there are 2 identical cab companies. For each, the arrival rate of customer is 4 per hour and average service time is 12 minutes, assuming poisons arrival an exponential service time distribution find out whether there will be a change in utilisation factor rho or waiting time in the queue WQ if the 2 company is merged. You see there are 2 separate cab companies, right, customers call to each at 4 per hour and their service time is 12 minutes. So what happens when they operate individually? And what happens when they pool their resources and you know served as one company?

(Refer Slide Time: 26:42)



Another Example

In a small city, there are two identical cab companies. For each, Arrival rate of customers is 4 per hour and Average Service Time is 12 minutes. Assuming Poisson arrival and exponential service time distribution, find out whether there will be change in the utilization factor ρ or Waiting time in queue Wq, if the two companies merge.

Answer: Without pooling of resources at any one company

For each company, Average arrival rate λ = 4 per hour.

Average service time 1/μ = 12 minutes, and hence μ = 5 per hour.

Hence, the utilization factor, ρ = λ/μ = 4/5 = 0.80

Average number of customers in the system,

$Lq = \rho^2/(1 - \rho) = (4/5)^2/(1 - 4/5) = 16/5$

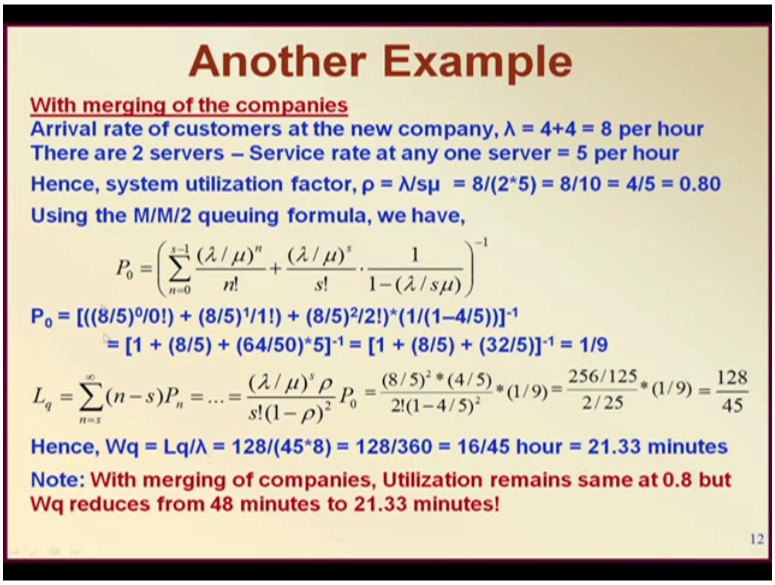
Average waiting time in the queue for the customers,

Lq/λ = (16/5)/4 = 16/20 hour = 4/5 hour = 48 minutes.

11

So the first thing is rather easy that without pulling the lambda is 4, Mu is 5, the utilisation is 80 percent and LQ rho square by 1 minus rho is 16 by 5 and therefore average waiting time WQ comes out to be LQ by lambda is 16 by 20 equal to 48 minutes, right. So look here without pulling it takes 48 minutes for a person to wait in the queue before the person gets a cab service.

(Refer Slide Time: 27:23)



Another Example

With merging of the companies

Arrival rate of customers at the new company, λ = 4+4 = 8 per hour

There are 2 servers – Service rate at any one server = 5 per hour

Hence, system utilization factor, ρ = λ/sμ = 8/(2*5) = 8/10 = 4/5 = 0.80

Using the M/M/2 queuing formula, we have,

$$P_0 = \left(\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \cdot \frac{1}{1-(\lambda/s\mu)} \right)^{-1}$$

$P_0 = [((8/5)^0/0!) + (8/5)^1/1!) + (8/5)^2/2!)*(1/(1–4/5))]^{-1}$

$= [1 + (8/5) + (64/50)*5]^{-1} = [1 + (8/5) + (32/5)]^{-1} = 1/9$

$$L_q = \sum_{n=s}^{\infty}(n-s)P_n = \ldots = \frac{(\lambda/\mu)^s \rho}{s!(1-\rho)^2} P_0 = \frac{(8/5)^2 * (4/5)}{2!(1-4/5)^2} * (1/9) = \frac{256/125}{2/25} * (1/9) = \frac{128}{45}$$

Hence, Wq = Lq/λ = 128/(45*8) = 128/360 = 16/45 hour = 21.33 minutes

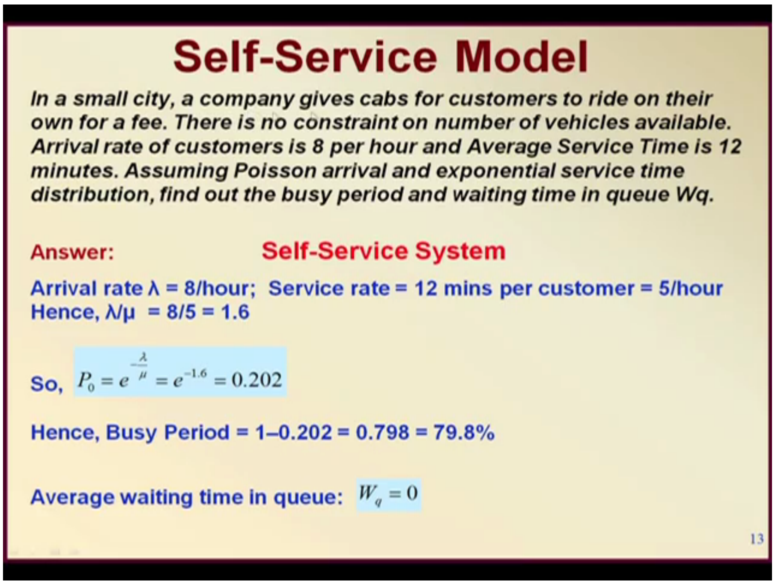
Note: With merging of companies, Utilization remains same at 0.8 but Wq reduces from 48 minutes to 21.33 minutes!

12

What happens if we pool the resources? When you pool the resources then lambda becomes 8, service time becomes 5, the system utilisation factor is lambda by s Mu becomes 80 percent, using the MM 2 queuing formula you know lambda by Mu to the power n by factorial n, Lambda by Mu to the power s by factorial s and 1 by 1 minus rho inverse that inverse must remembered.

So when you compute this we get 1 by 9 and LQ formula shows, LQ is 128 by 45, so WQ will be 128 by 45, divided by 1 by 8, I mean 8, so gives 128 by 360 , right that is 16 by 45 hour or 21.33 minutes. So what happens you know utilisation at 80 percent, but WQ reduces to from 48 minutes to 21.33 minutes, so you know that is the kind of advantage one can get out of pooling of resources.

(Refer Slide Time: 28:38)



Self-Service Model

In a small city, a company gives cabs for customers to ride on their own for a fee. There is no constraint on number of vehicles available. Arrival rate of customers is 8 per hour and Average Service Time is 12 minutes. Assuming Poisson arrival and exponential service time distribution, find out the busy period and waiting time in queue Wq.

Answer: Self-Service System

Arrival rate λ = 8/hour; Service rate = 12 mins per customer = 5/hour
Hence, λ/μ = 8/5 = 1.6

So, $P_0 = e^{-\frac{\lambda}{\mu}} = e^{-1.6} = 0.202$

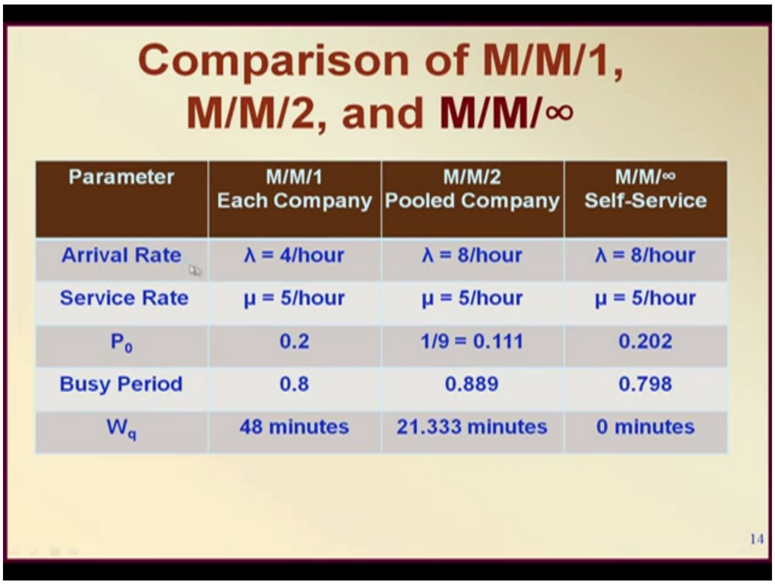
Hence, Busy Period = 1–0.202 = 0.798 = 79.8%

Average waiting time in queue: $W_q = 0$

13

And what happens in self-service, the Wq is 0 because nobody waits. Look how the problem changes, the company gives cabs for customer to ride on their own for a fee and but then is no constraint on number of vehicles available that is the constraint, so really it is not really possible to have and infinite self-service facility. Arrival rate of customer is 8 per hour and average service time is 12 minutes, so what will happen? What will really happen is the P 0 is 20.2 percent and the busy period becomes 79.8 percent and average waiting time is 0, right.

(Refer Slide Time: 29:16)



Comparison of M/M/1, M/M/2, and M/M/∞

Parameter	M/M/1 Each Company	M/M/2 Pooled Company	M/M/∞ Self-Service
Arrival Rate	λ = 4/hour	λ = 8/hour	λ = 8/hour
Service Rate	μ = 5/hour	μ = 5/hour	μ = 5/hour
P_0	0.2	1/9 = 0.111	0.202
Busy Period	0.8	0.889	0.798
W_q	48 minutes	21.333 minutes	0 minutes

14

So how do they compare? The arrival rate 4 per hour, 8 per hour, 8 per hour, service rate 5 per hour in all the 3 cases, P 0 20 percent, 11.1 percent and 20.2 percent. Busy period 80 percent, 88.9 percent, 79.8 percent, WQ 48 minutes, 21.333 minutes and 0 minutes, so that is really shows a comparison that if the 2 companies really merge and really serve together they can definitely give a much better service.

Because when one is free the other may be you know utilized, so they can take the load of the other one during their free period and that is how the pooling of resources happens and that is why you know if you really improve queuing system it is imperative that we pool the resources and tried to become, try to make multiple server system out of several single server systems. Thank you very much.