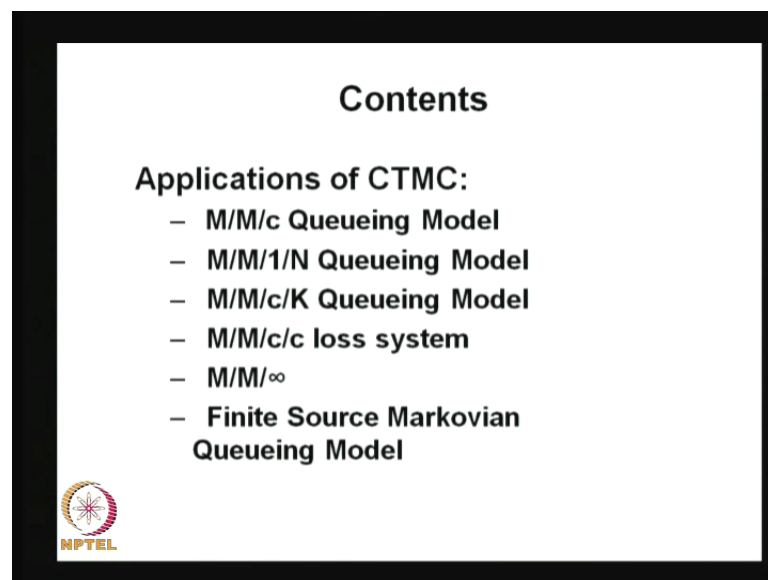


**Stochastic Processes**  
**Prof. Dr. S. Dharmaraja**  
**Department of Mathematics**  
**Indian Institute of Technology, Delhi**

**Module - 5**  
**Continuous-time Markov Chain**  
**Lecture - 5**  
**Simple Markovian Queueing Models**

This is the lecture five application of a continuous-time Markov chain in simple Markovian queueing models. The first lecture we have discussed the definition of a stochastic process in particular continuous-time Markov chain, then we have considered the Kolmogorov differential equation Chapman-Kolmogorov equations the transient solutions for the CTMC. In the second lecture, we have discussed the special case of continuous-time Markov chain that is birth-death process we have discussed in lecture two. Lecture three the special case of birth-death process it is a very important stochastic process that is Poisson process is discussed in the lecture three.

(Refer Slide Time: 02:01)

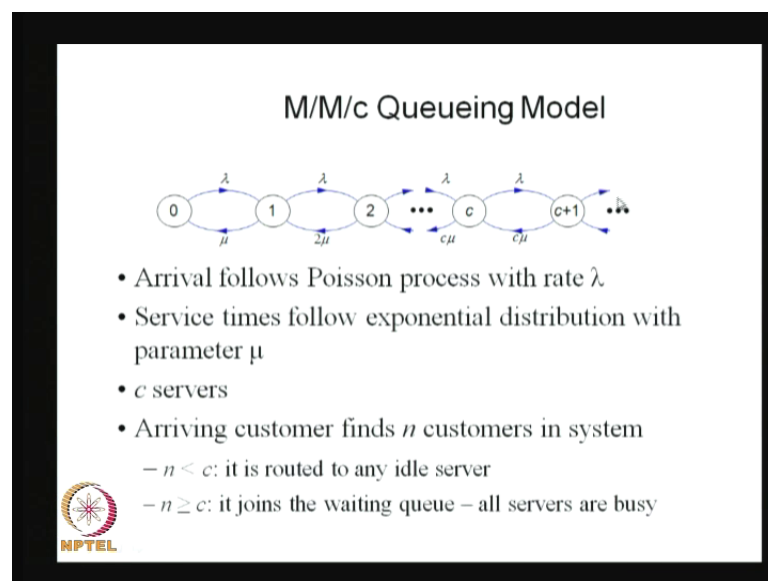


In the lecture four we have discussed the M/M/1 queueing model, that is a very special and important queueing model and the underlying stochastic process for the M/M/1 queueing model that is a birth-death process with birth rates are  $\lambda$ , and death rates are  $\mu$ . In the fourth lecture we have discussed only the M/M/1 queueing model. In this

lecture we are going to consider the other simple Markovian queuing models as an application of a continuous-time Markov chain.

So, in this lecture I am planning to discuss other than M/M/1 queuing model I am going to discuss the simple Markovian queuing models starting with the M/M/c infinity queuing model, then the finite capacity model Markovian set of M/M/1/N queuing model. Then I am going to discuss the multi server finite capacity model that is M/M/c/K queuing model. After that I am going to discuss the loss system that is M/M/c/c model, for infinite server model that is M/M/infinity also I am going to discuss. At the end I am going to discuss the finite source Markovian queuing model also, whereas the other five models the population is infinite source. So, the last one is a finite source Markovian queuing model also I am going to discuss as the application of continuous-time Markov chain.

(Refer Slide Time: 03:03)



The first model is a multi server infinite capacity Markovian queuing model. The letter M denotes the inter arrival time is exponentially distributed with parameter  $\lambda$ . The service time by the each server that is exponentially distributed with the parameter  $\mu$  and all we have more than one server. Suppose you consider as a  $c$  where  $c$  is a positive integer, and all the servers are identical, and each server is doing the service which is exponentially distributed with the parameter  $\mu$  which is independent of the all other server, and the service time is independent with the inter arrival time also.

With this assumptions if you make a random variable  $x$  of  $t$  is the number of customers in the system at any time  $t$  that is a stochastic process. Since the possible values of number of customers in the system at any time  $t$  that is going to be 0, 1, 2 and so on; therefore, it is a discrete state and you are observing the queuing system at any time  $t$  therefore, it is a continuous time, so discrete state continuous times stochastic process, and if you observe the system keeps moving into the different states because of either arrival or the service completion from the any one of the  $c$  servers.

So, suppose there is no customer in the system, the system moves from the state 0 to 1 by one arrival. So, the inter arrival time is exponentially distributed; therefore, the rate in which the system is moving from the state 0 to 1 is  $\lambda$  like that you can visualize the rates for system moving from 1 to 2, 2 to 3 and so on, whereas whenever the system size is 1, 2 and so on till  $c$ , since we have a  $c$  number of service in the system whoever is entering into the system they will start getting the service immediately. Suppose the system goes from state 1 to 0; that means the customer enter to the system and he get the service immediately and the service time exponentially distributed with the parameter  $\mu$ . Therefore, whenever the service is completed the system goes from the state 1 to 0; therefore, the rate is  $\mu$ , whereas from 2 to 1 there are two customers in the system and both are under service; at any time if any one of the servers completes the service then the system moves from 2 to 1.

So, the service completion will be minimum of the service time of both the servers. Since each server is doing the service exponentially distributed with the parameter  $\mu$ ; therefore, the minimum of two exponential and both are independent also, therefore, that is also going to be exponentially distributed with the sum of parameters. So, it is going to be parameter will be  $\mu$  plus  $\mu$  that is  $2\mu$ . So, the system moves from the state 2 to 1 will be the rate will be  $2\mu$ , like that it will keep going till the state from  $c$  to  $c - 1$ ; that means we have  $c$  servers. Therefore, whenever the system size is also less than or equal to  $c$  that means all the customers are under service.

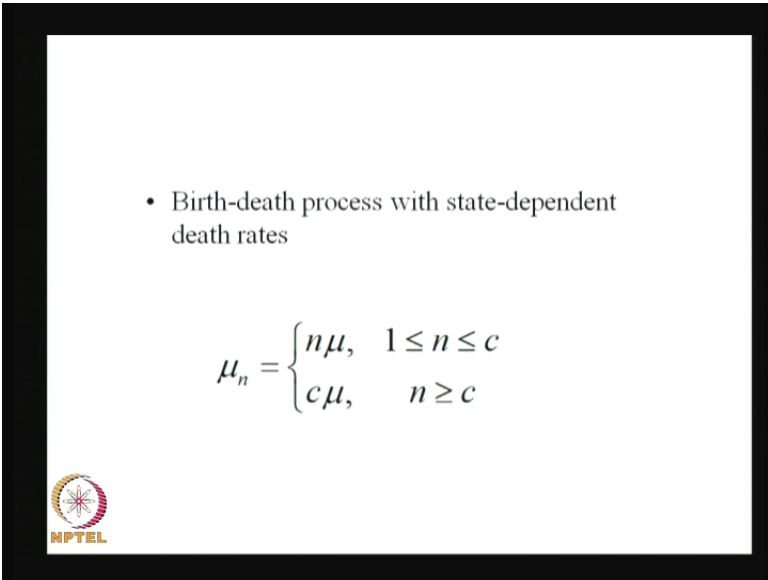
Now we will discuss the rate in which system is moving from the state  $c + 1$  to  $c$ . The system state is the  $c + 1$ ; that means when the number of customer in the system that is  $c + 1$ , we have  $c$  servers; therefore, one customer will be waiting for the service waiting in the queue. Therefore, the system is moving from  $c + 1$  to  $c$  that is nothing but one of the server has completed the service out of  $c$  servers; therefore, the rate will be

the completion service time will be exponential distribution with the parameter  $c\mu$ , not  $c + 1\mu$  it is we have only  $c$  servers.


Therefore the minimum of exponentially distributed with the parameters  $\mu$  and so on with the  $c$  exponentially distributed random variables; therefore, that is going to be exponential distribution with the parameter  $\mu$  plus  $\mu$  plus, there are  $c\mu$ 's; therefore, it is going to be  $c\mu$ . Like that the rate will be the death rate will be  $c\mu$  after  $c + 1$  onwards whereas from 0 to  $c$  it will be  $\mu$ ,  $2\mu$ ,  $3\mu$  and so on till  $c\mu$  after that it will be  $c\mu$  from the state from  $c + 1$  to  $c$ ,  $c + 2$  to  $c + 1$  and so on.

And if you see the state transition diagram you can observe that this is a birth-death process. So, before that let me explain what is M/M/c infinity means whenever  $c$  customers or  $c$  servers are any one of the  $c$  servers are available then the customers will get the service immediately. If all the  $c$  servers are busy then the customer has to wait till any one of the  $c$  servers is going to be completing their service. So, that is the way the system works; therefore, you will have the system size.

(Refer Slide Time: 09:43)



- Birth-death process with state-dependent death rates

$$\mu_n = \begin{cases} n\mu, & 1 \leq n \leq c \\ c\mu, & n \geq c \end{cases}$$


The system size the underline stochastic process is going to be a birth-death process. It is a special case of continuous-time Markov chain because the transitions are only the neighbor's transition with the forward rate that is  $\lambda$ , and backward rates are the death rates are going to be  $\mu$ ,  $2\mu$  and so on. Therefore, this is the special case of a continuous-time Markov chain; the underlying stochastic process for the M/M/c infinity

model that is a birth-death process. The birth rates are  $\lambda$  whereas the death rates depends on the  $n$  the  $\mu_n$  is the function of  $n$ ; therefore, it is called a state-dependent death rates. It need not be the function  $n$  times  $\mu$  it can be a function of  $n$  then we can use the word state-dependent.

So, here it is a linear function. So, state-dependent death rates and the death rates are  $n$  times  $\mu$  whenever  $n$  lies between 1 to  $c$  and the  $\mu_n$  is going to be  $c$  times  $\mu$  for  $n$  is greater than or equal to  $c$ ; that you can observe it from the state transition diagram also; the death rates are going to be  $c\mu$  here also  $c\mu$  and so on. Therefore, this is the birth-death process with the state-dependent death rates.


(Refer Slide Time: 11:13)

### M/M/c Queuing Model

- Steady-state or equilibrium solution when  $\frac{\lambda}{c\mu} < 1$

$$p_n = \begin{cases} \frac{\lambda^n}{n! \mu^n} p_0 & 1 \leq n \leq c \\ \frac{\lambda^n}{c^{n-c} c! \mu^n} p_0 & n > c \end{cases}$$

Using normalizing constant

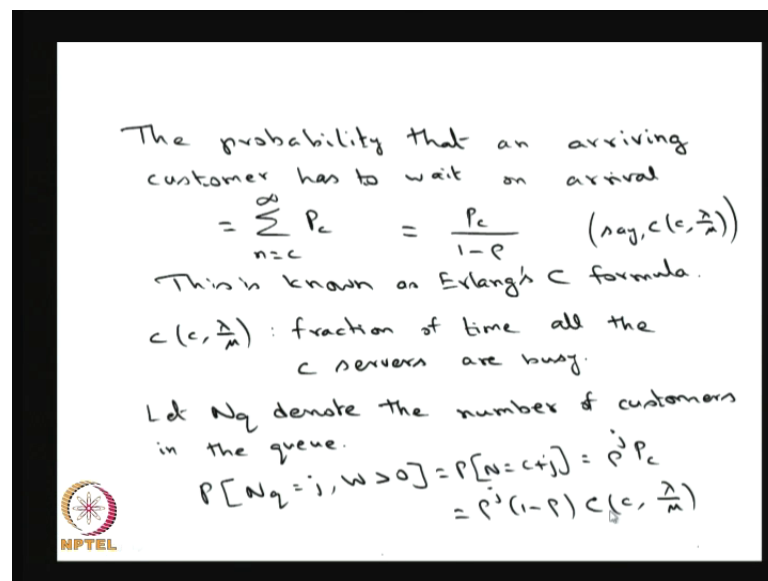
$$\sum_{n=0}^{\infty} p_n = 1 \Rightarrow p_0 = \left[ \sum_{n=0}^{c-1} \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n + \frac{1}{c!} \left( \frac{\lambda}{\mu} \right)^c \left( \frac{c\mu}{c\mu - \lambda} \right) \right]^{-1}$$


Now our interest is to find out the steady-state or equilibrium solution. Since it is infinite capacity models if you observe the birth-death process with the infinite state space then you need condition, so that the steady state probability exists. So, whenever  $\lambda$  by  $c\mu$  is less than 1 you can find out the limiting probabilities. So, sometimes I use the letter  $P_n$ , sometimes I use the word  $p_n$ , both are one and the same. So, you find out the steady-state probability by solving  $P_q$  is equal to 0, and the summation of  $p_i$  is equal to 1. And if you recall the birth-death process of steady state probabilities the  $p_0$  has the one divided by the series; whenever the denominator series converges then you will get the  $P_n$ 's.

So either I use  $P_n$ 's or  $p_n$ 's both are one and the same. So, here summation of  $P_i$  is equal to 1 and if you make a vector  $p$ ,  $p$  times the  $q$  is equal to 0 if you solve that equation, and the denominator of  $P_0$  that expression that is going to be converged only if  $\lambda / c\mu$  is less than 1. So therefore, whenever this condition is there the queuing system is stable also. If you put  $c$  is equal to 1 you will get the M/M/1/q. So, using the normalizing condition you are getting the  $p_0$  and  $P_0$  is 1 divided by this. So, this is the series. So, this series is going to be converged only if this condition is satisfied.

So, by solving that equation you are getting  $P_n$  in terms of  $P_0$  and using normalizing constant you are getting a  $P_0$ ; therefore, this is a steady-state also known as the equilibrium solution for the M/M/c infinity model. So, here we are using the birth-death process with the birth rates are  $\lambda$  and the death rates are given in this form and use the same logic of the stationary distribution for the birth-death process; using that we are getting the steady-state or equilibrium solutions for the M/M/c model.

(Refer Slide Time: 13:43)



The probability that an arriving customer has to wait on arrival


$$= \sum_{n=c}^{\infty} P_n = \frac{P_c}{1-\rho} \quad \left( \rho, c, \frac{\lambda}{\mu} \right)$$

This is known as Erlang's C formula.

$C(c, \frac{\lambda}{\mu})$  : fraction of time all the  $c$  servers are busy.

Let  $N_q$  denote the number of customers in the queue.

$$P[N_q \geq 1, W > 0] = P[N \geq c] = \rho^c P_c$$

$$= \rho^c (1-\rho) C(c, \frac{\lambda}{\mu})$$


Other than the steady-state probability we can get some more measures. The first one is the probability that the arriving customer has to wait an arrival. What is the probability that the arriving customer has to wait on arrival? So, that means the number of customers in the system is greater than or equal to  $c$ , then only the customer has to wait. So, the probability you add the probability of  $p_c$ , sorry  $p_n$ , sorry here it is a mistake  $p$  suffix  $n$

where  $n$  is running from  $c$  to infinity, if you add all those probabilities that is going to be  $\rho^c$  divided by  $1 - \rho$ , and this probability is known as the Erlang's C formula for a multi server infinite capacity model; that too I am denoting with the letter  $C$  of  $c$  comma  $\lambda$  by  $\mu$

Because you need number of servers in the system and you need  $\lambda$  as well as  $\mu$ ; if I know this quantity I can find out what is the Erlang's C formula. This is a very important formula; using that you can find out what is the optimal  $c$  such a way that the probability has to be minimum. You can find out what is optimal number of servers needed to have some upper bound probability of arriving customer has to wait; therefore, the Erlang's C formula is very useful in performance analyses of any system.

The next quantity is  $N_q$  denotes the number of customers in the queue. So, either I use the letter  $N$  suffix  $q$ ; earlier I used the letter  $q$  itself. So, for that I am finding the joint distribution of what is the probability that the number of customers in the queue is  $j$  and the waiting time is going to be greater than zero,  $W$  is used for the waiting time. So, the waiting time is going to be greater than zero. That is same as the number of customers in the system that is  $c$  plus  $j$ . What is the probability that  $j$  customers in the queue as well as the waiting time is greater than zero; that is same as what is the probability that  $c$  plus  $j$  customer in the system. We will do the little simplification so you will get this joint probability in terms of Erlang's C formula.

(Refer Slide Time: 16:30)

Thus,


$$P\{N_q = j / W > 0\} = \frac{P\{N_q = j, W > 0\}}{P\{W > 0\}}$$

$$= (1 - \rho) \rho^j, \quad j = 0, 1, \dots$$

Expected number of busy servers

$$E(B) = \sum_{n=0}^{c-1} n P_n + \sum_{n=c}^{\infty} c P_n = c \rho$$

Expected number of idle servers

$$E(I) = E(c - B) = c - c \rho = c(1 - \rho)$$


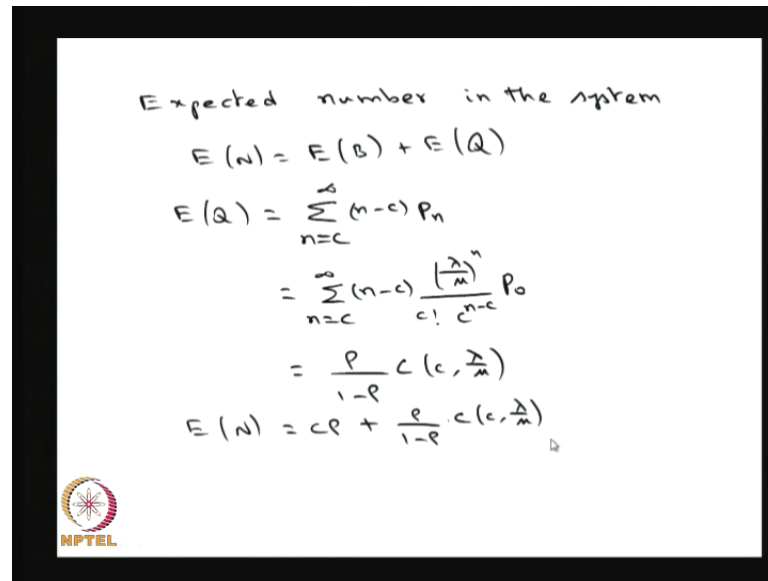

So, using that I am finding the conditional probability, what is the conditional probability that  $j$  customer in the queue given that the waiting time is greater than zero. If I do little simplification I will get  $1 - \rho$  times  $\rho$  power  $j$  where  $\rho$  is  $\lambda$  divided by  $c\mu$ . This is nothing but the probability mass function of a geometric distribution. This is the probability mass function of a geometric distribution; therefore, this conditional probability is geometrically distributed with the parameter  $\rho$ . From this we can find out the next measure is expected number of busy service. What is the average number of busy servers? That is nothing but the summation of  $n$  equal to zero to  $c - 1$  of  $n$  times  $p^n$ .

That means whenever the system size is less than  $c$  only those many servers are busy and with the probability. Whenever  $n$  customers or more than  $n$  customers in the system all the  $c$  servers are going to be busy; therefore,  $c$  times  $p$ . If you simplify you will get  $c$  times  $\rho$ . So, that is the expected number of busy servers. Once I know the expected number of busy servers I can find out what is the expected number of idle servers also, it is a negation, that is expected number of idle server is nothing but expectation of it is a random variable.

So, idle number is nothing but there are totally  $c$  servers in the system; therefore,  $c$  minus busy servers are capital  $B$ . Therefore,  $c$  minus  $B$  is same as  $i$ . So, the expectations satisfies the linear property; therefore, expectation of  $i$  is same as expectation of  $c$  minus  $B$ .  $c$  is a constant, and  $B$  is a random variable; therefore, it is a  $c$  minus expectation of a  $B$ . expectation of  $B$  just now we got  $c$  times  $\rho$ . Therefore, the expected number of idle server is  $c$  times  $1 - \rho$ . So, other than stationary distribution for the  $M/M/c$  model we are getting what is the probability that arriving customer has to wait, and we are getting the conditional probability of  $j$  customers in the queue given that waiting time is greater than zero as well as this expected quantities we are getting.



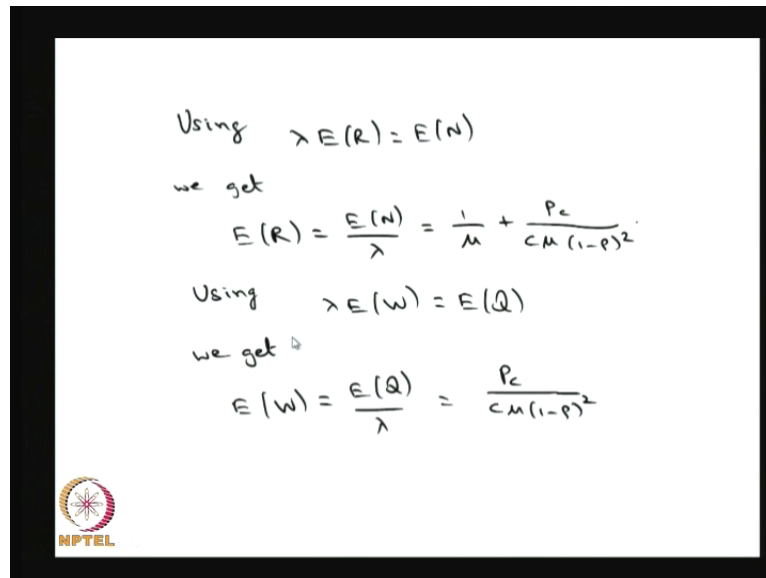
(Refer Slide Time: 19:10)


$$\begin{aligned} &\text{Expected number in the system} \\ E(N) &= E(B) + E(Q) \\ E(Q) &= \sum_{n=c}^{\infty} (n-c) P_n \\ &= \sum_{n=c}^{\infty} (n-c) \frac{(\frac{\lambda}{\mu})^n}{c! c^{n-c}} P_0 \\ &= \frac{\rho}{1-\rho} C(c, \frac{\lambda}{\mu}) \\ E(N) &= c\rho + \frac{\rho}{1-\rho} C(c, \frac{\lambda}{\mu}) \end{aligned}$$


Also we can find out what is the expected number of customers in the system; that is nothing but expected number is nothing but expected of the busy servers plus expected number in the queue. Earlier I used the notation n suffix queue; n suffix queue and queue are both one and the same. So, I can compute what is the expectation of queue come and do the little simplification. Then I can substitute expectation of queue here; therefore, I will get expected number of customers in the system that involves the Erlang's C formula. So, this Erlang's C formula is used to get the expected number of customers in the system and then later we can do some optimization over the probability expected number with the specified c and lambda by mu.

So, using Little's formula I can find out the expected time spent in the system, because I know what is the arrival rate and from the stationary distribution I got expected number in the system in a steady-state. Therefore, since I know lambda and expectation of n I can get expectation of r where r is the response time or so joint time is total time spent in the system. So, that expectation is going to be expectation of n divided by lambda; do little simplification, you will get expectation of r.


(Refer Slide Time: 20:12)



Using  $\lambda E(R) = E(N)$   
 we get  

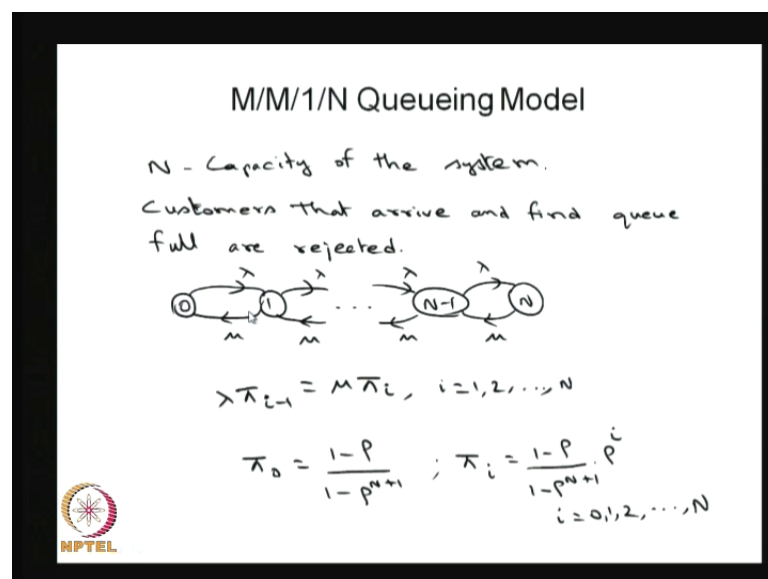
$$E(R) = \frac{E(N)}{\lambda} = \frac{1}{\mu} + \frac{\rho_c}{c\mu(1-\rho)^2}$$
 Using  $\lambda E(W) = E(Q)$   
 we get  

$$E(W) = \frac{E(Q)}{\lambda} = \frac{\rho_c}{c\mu(1-\rho)^2}$$



You can apply the Little's formula in the Q level also. So, this is the system level and you can apply the Q level also. So, here lambda times the expectation of waiting time is same as expectation of number of customer in the queue. So, expectation of waiting time or average time is same as expectation of a Q divided by lambda. So, since the M/M/c infinity Q the underlying stochastic process is the birth-death process; therefore, we are getting the all the measures using the birth-death logic.

(Refer Slide Time: 21:42)



Next I am going for the finite capacity. So, the  $N$  is the capacity of the system; that means whenever the customers arrives and find the queue full that customer will be rejected. Therefore, at any time the number of customers in the system if you make it as a random variable and that random variable takes the possible values from zero to capital  $N$ ; therefore, the state space is finite. The number of customers in the system makes any time  $t$ , that is a random variable, and you will have a stochastic process. And since the inter arrival time is exponentially distributed services exponential distributed only one server finite capacity; therefore, the underlying stochastic process is a birth-death process with the birth rates  $\lambda$  and the death rates  $\mu$ .

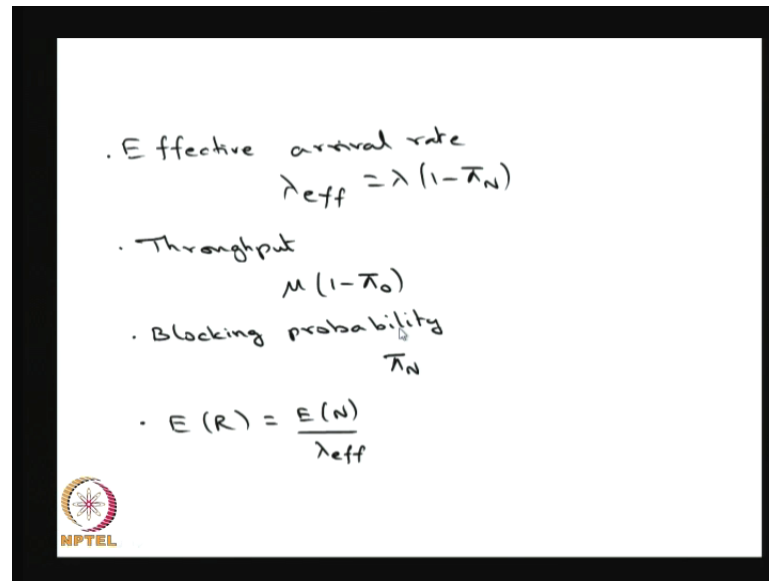
If you see the queue matrix for a this one infinitesimal generator matrix that is a tridiagonal matrix with all the half diagonals are  $\lambda$ s as well as  $\mu$  and the diagonals are minus  $\lambda$  plus  $\mu$  except the first term and the last term except the first row and last row. Our interest is to get the stationary distribution; later I going to explain the time-dependent solution also. So, to get the stationary distribution either you write  $\pi Q$  is equal to zero and the summation of  $\pi$  is equal to one and solve that, or you write the balance equation. The  $\pi Q$  is equal to zero that will land up a balance equation. So, some books write this as a balance equation, what is the inflow rate and what is the outflow rate? Both are going to be same whenever the system reaches the equilibrium solutions equilibrium state.

Therefore, the outflow is  $\lambda$  times this; the inflow is  $\mu$  times  $\lambda$  one. Like that you can go for understanding the balance equation for this state and second and so on. This also satisfies the time, this is also called satisfying the time reversible equation; therefore, one can use the time reversible property of birth-death process. So, you can find out  $\pi_i$ 's easily using the time reversible equation itself. You do not want to use a  $\pi Q$  is equal to zero; instead of that you can write the time reversible equation since it is satisfied by all the states. Now you can use the summation of  $\pi_i$  is equal to 1  $i$  starting from 1 to  $n$ ; therefore, you will get  $\pi_0$  naught, and here the birth-death process with the finite state space.

Therefore, the  $\pi_0$  naught will be 1 divided by the denominator series that is a finite series finite terms in it. Therefore it always converges immaterial of the value of  $\lambda$  and  $\mu$ ; therefore, you will get  $\pi_0$  naught without any restriction over  $\lambda$  and  $\mu$ . So, once you get the  $\pi_0$  naught you can get  $\pi_i$ 's in terms of  $\pi_0$  naught. Therefore, that is 1

minus rho divided by 1 minus rho power n plus 1 times rho power i where rho is lambda by mu. So, this is the birth; the underline stochastic process is the birth-death process with the birth rates lambda and death rates mu. So, you can use all the concepts of the birth-death process and you can analyze the system in an easy way. So, this is a steady-state probability.

(Refer Slide Time: 25:43)



Handwritten notes on a slide:

- Effective arrival rate  

$$\lambda_{eff} = \lambda (1 - \pi_N)$$
- Throughput  

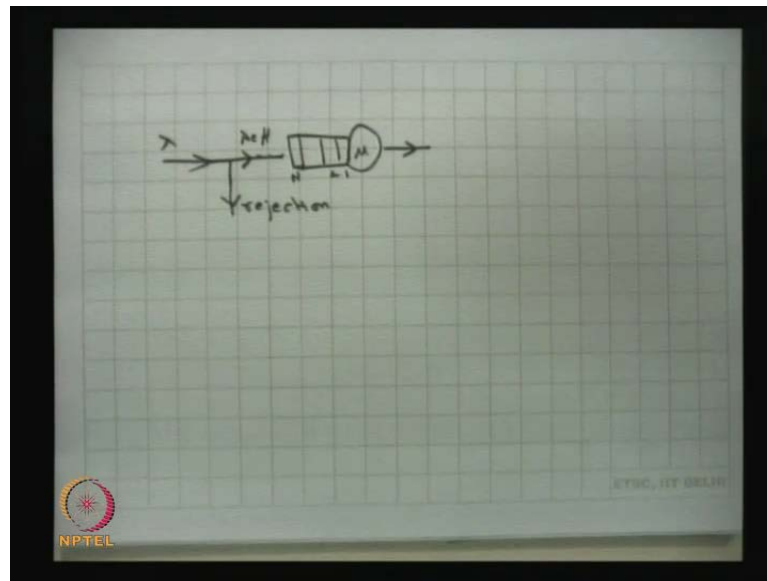
$$\mu (1 - \pi_0)$$
- Blocking probability  

$$\pi_N$$
- $E(R) = \frac{E(N)}{\lambda_{eff}}$

NPTel logo is visible in the bottom left corner of the slide.

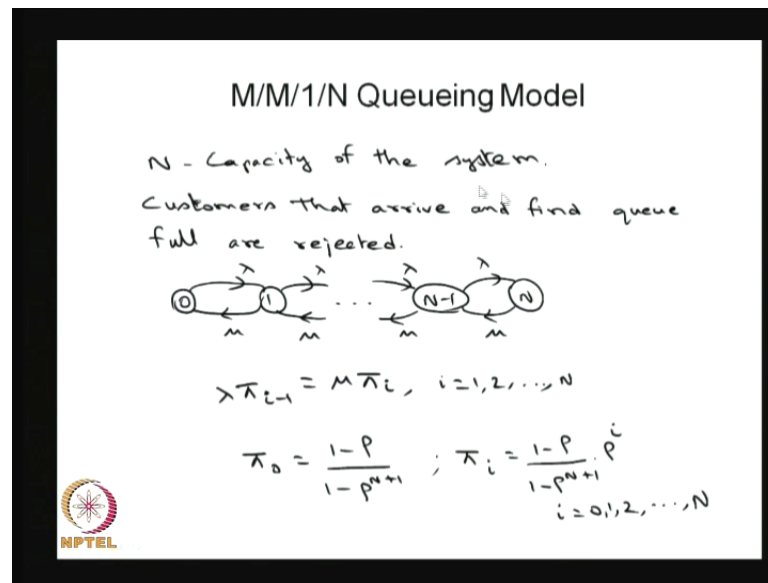
Once you know the steady-state probability you can get the other measures also. Here the other important thing is called the effective arrival rate; that means the system the queuing system is a finite capacity.

(Refer Slide Time: 26:02)



So, maximum  $N$  customers can wait in this system and the service rate is  $\mu$ , the arrival rate is  $\lambda$  from the infinite population. So, whenever the system size is full the customer is rejected; therefore, there is a rejection. After the service is completed the system leaves the system. So, the effective arrival rate is nothing but what is the rate in which the customers are entering into the system. So, there is a partition here. So, the effective arrival rate is  $\lambda H$ . That rate will be, what is the probability that the system is not full multiplied by the arrival rate  $\lambda$  that is going to be the  $\lambda$  effective. Whenever the system is not full that proportion of the time or the probability is  $1 - \pi_N$  where  $\pi_N$  is the steady state probability just now we got it.

(Refer Slide Time: 27:20)



From here you can get  $\pi_N$ ; that is the probability that the system is full and  $1 - \pi_N$  is the probability that the system is not full and multiplied by the arrival rate that is going to be the  $\lambda$  effective.

And you can also find out the throughput. Throughput is nothing but what is the rate in which the customers are served per unit of time. The service rate is  $\mu$ , and this is the probability that the system is not empty  $1 - \pi_0$ . Therefore  $1 - \pi_0$  times  $\mu$  that is the rate in which the customers are served in the M/M/1/N system. Whenever the system is not empty multiplied by that probability multiplied by  $\mu$  that is going to be the throughput. By using the time-reversible equation the  $\mu$  times  $1 - \pi_0$  you can get in terms of  $\lambda$  equivalent also, but throughput is the service rate multiplied by what is the probability that the system is not empty.


Since it is a finite capacity system one can find out the blocking probability also. Blocking probability is nothing but the probability that the customers are blocked. The customers are blocked whenever the system is full; therefore, the blocking probability same as the probability that the system is full that is  $\pi_N$ . Once you know the steady-state probabilities you can find out the average number of customers in the system, and using the Little's formula you can get expected time spent in the system by any customer divided by not  $\lambda$  it is  $\lambda$  effective because the effective arrival rate is used in the Little's formula not the arrival rate.

For M/M/1 infinity system the effective arrival rate and arrival rate are one and the same, because there is no blocking; therefore, the probability of 1 minus  $\rho$  that is equal to 1 only. Therefore, the effective arrival rate and the arrival rate are same for an infinite capacity system, because there is no blocking. For a finite capacity system the effective arrival rate has to be computed. Similarly, we have to go for finding the M/M lambda effective for the M/M/c/K model also. So, other than the stationary distribution or equilibrium probabilities we are getting the other performance measures using the birth-death process concepts.

(Refer Slide Time: 30:10)

### M/M/c/K Queueing Model

- Arrival follows Poisson process with rate  $\lambda$
- Service times follow exponential distribution with parameter  $\mu$
- $c$  Servers with system capacity  $K$
- Arriving customer find  $n$  customers already in system, where, if
  - $n < c$ : it is routed to an idle server
  - $n \geq c$ : it joins the waiting queue – all servers are busy
- Customers forced to leave the system if already  $K$  present in the system.



Now I am moving into M/M/c/K model queueing model. So, here the change is instead of one server in the M/M/1 model you have more than one server's  $c$ , and you have finite capacity that is capital  $K$  capacity of the system. So, the arrival follows a Poisson process; service is exponential. We have  $c$  identical servers, the capacity is capital  $K$ . And this is the scenario in which whenever the system size is less than  $c$  it will be routed into the idle server. If it is greater than or equal to  $c$  that means all the servers are busy; that means the customer has to wait.

But if the system size is full that means  $c$  customers are under service and  $k$  minus  $c$  customers are waiting in the queue for the service. Then whoever comes, it will be rejected, forced to leave the system. Therefore, you have the waiting as well as blocking, because it is a finite capacity there is a blocking. And since you have always we choose

the K such that it is K is always greater than or equal to c. If K is equal to c then it is a loss system. If the K is greater than c then K minus c customers maximum can wait in the system in the queue.

(Refer Slide Time: 31:47)


### M/M/c/K Queueing Model

- Birth death process with state dependent death rates

$$\mu_n = \begin{cases} n\mu & , 1 \leq n < c \\ c\mu & , c \leq n \leq K \end{cases}$$

- Steady-state or equilibrium solution

$$\pi_n = \begin{cases} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n \pi_0 & , 0 \leq n < c \\ \frac{1}{c!} \left(\frac{\lambda}{\mu}\right)^n \pi_0 & , c \leq n \leq K \end{cases}$$



Therefore, the underlining stochastic process, here the stochastic process is again number of customers in the system at any time t. Therefore, this stochastic process is also going to be a continuous time Markov chain. Because of this assumptions inter arrivals are exponential distributions are with each service by each server is exponentially distributed and all are independent and so on. So, with these assumptions this stochastic process is a continuous time Markov chain, and at any time only one forward or only one backward the system can move.

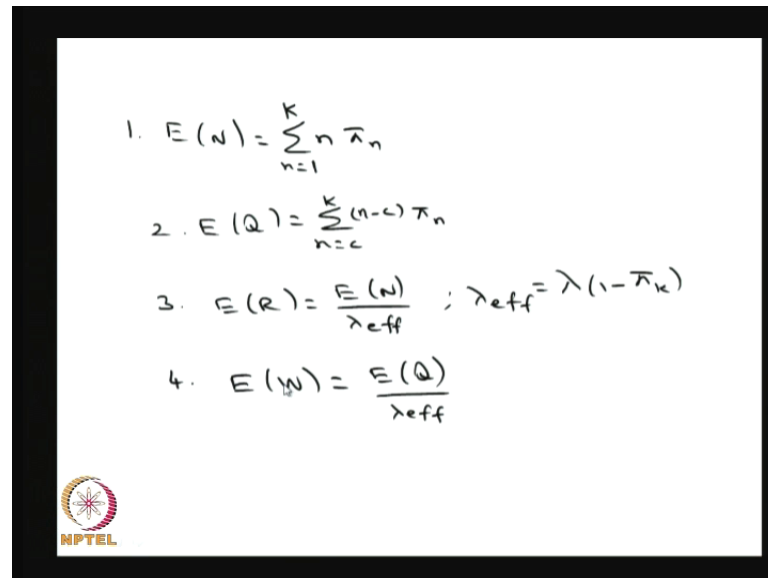
Therefore, it is going to be a birth-death process also, and the birth rates are lambda, because it is an infinite source population. So, all the lambda N's are going to be lambda whereas the death rates are state-dependent; that is going to be n times mu lies between 1 to c, from c to k onwards it is going to be c mu. So, I have not drawn the state transition diagram for M/M/c/K, but you can visualize the way we have M/M/1/n and M/M/c model with the combination of that; that is going to be the state transition diagram.

Since it is a finite capacity model it is easy to get the steady state and equilibrium the solution. So, first you solve pi Q is equal to zero; that means you write pi n's in terms of pi naught, and use the normalizing constant summation of pi i is equal to 1, using that



you will get  $\pi_n$ . So, I have not written here. So, use the normalizing constant the summation of  $\pi_i$  equal to 1 and get the  $\pi_n$ , then substitute  $\pi_n$ ; therefore, you will get  $\pi_n$  in terms of  $\pi_0$  completely.

(Refer Slide Time: 33:48)



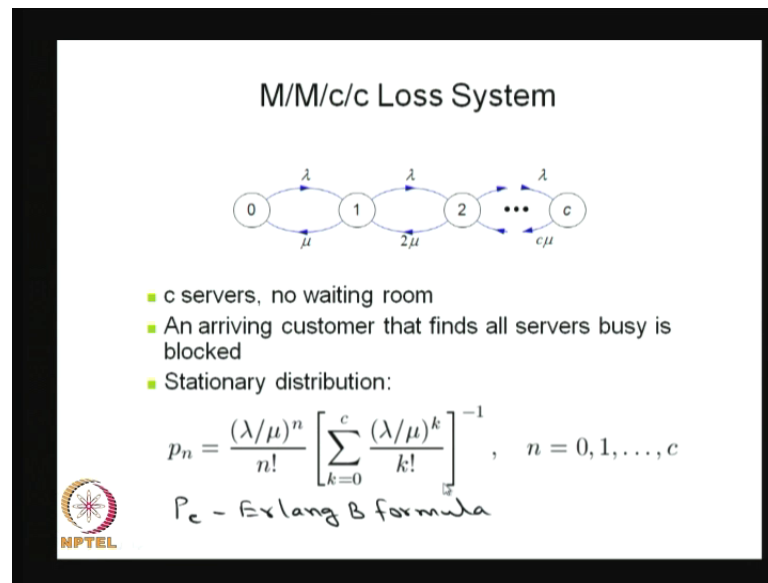
Handwritten mathematical formulas for M/M/1/N queue performance measures:

1.  $E(N) = \sum_{n=1}^K n \pi_n$
2.  $E(Q) = \sum_{n=c}^K (n-c) \pi_n$
3.  $E(R) = \frac{E(N)}{\lambda_{eff}} ; \lambda_{eff} = \lambda(1 - \pi_K)$
4.  $E(W) = \frac{E(Q)}{\lambda_{eff}}$

The NPTEL logo is visible in the bottom left corner of the slide.

After that you can get all the other average measures the way I have explained the M/M/1/N and the M/M/c infinity; the combination of that you can get the average number of customers in the system, average number of customer in the queue that is  $n$  minus  $c$  times  $\pi_n$ ; the summation goes from  $c$  to  $K$ , and the average time spent in the system. Since, it is a finite capacity you have to find out the lambda effective, effective arrival rate that is  $1$  minus its capacity is capital  $K$ ; therefore,  $1$  minus  $\pi_K$  and that is the probability that the system is not full. So, the effective arrival rate is  $\lambda$  times  $1$  minus  $\pi_K$ ; substitute here and get the average time spent in the system. And similarly, you can find out the average time spent in the queue also using the Little's formula.

(Refer Slide Time: 34:48)



Now I am moving into the fourth simple Markovian queuing model. First I started with the M/M/c infinity, M/M/1/N, then I did M/M/c/K; now I am going for capital K is equal to c that is loss system. It is not a queuing system, because we have a c servers, and the capacity of the system is also c. Example is you can think of a parking lot which has some c parking lots and the cars coming in to the system; that is if you make the assumption is inter arrival time is exponentially distributed and the car spending time in each parking lot that is exponentially distributed, then the parking lot problem can be visualized as the M/M/c loss system.

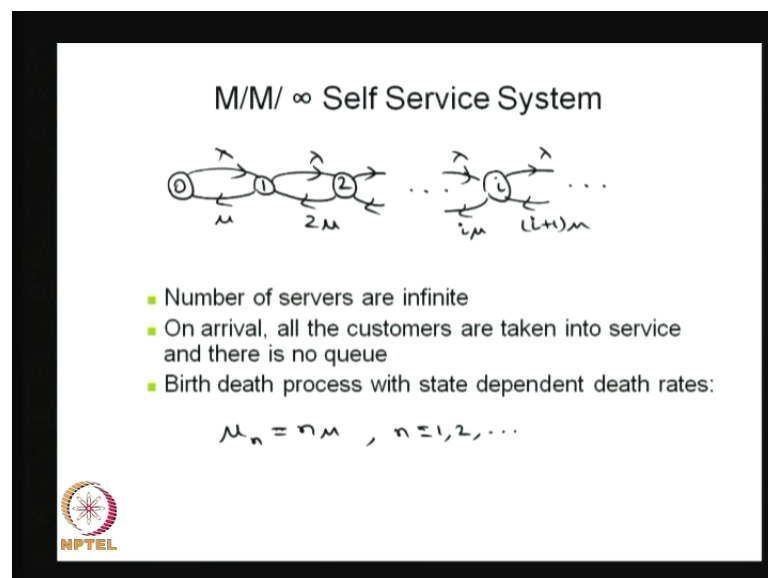
So, here we have a c identical servers and no waiting room. So, since it is a c capacity and the c waiting room you can think of a self service with the capacity c that also you can visualize. So, here the inter arrival times are exponentially distributed and the service by each server that is exponentially distributed with the parameter mu. Therefore, the system goes from 2 to 1, 1 to 0 and so on it is going to be how many customers in the system and completing their service; therefore, the time is exponentially distributed with the sum of those parameters accordingly.

Therefore, it is going to be 1 mu, 2 mu till c mu. Since, it is a finite capacity and so on it is an irreducible model positive recurrent; therefore, this steady-state probability exist, limiting probabilities also exist, and that is same as the equilibrium probabilities also.

Therefore, by using  $p_0$  is equal to  $1$  and the summation of  $p_i$  is equal to  $1$  you can get the steady-state or equilibrium probabilities that is  $p_n$ 's.

The  $p_c$  suffix that is nothing but the probability that the system is full and that is same known as the Erlang B formula. So, this is also useful to design the system for a given or what is the optimal  $c$  such a way that you can minimize the probability that the system is full. For that you need this formula; therefore, to do the optimization problem over the  $c$ . And here we denote a  $p_c$  that is the Erlang B formula, whereas, Erlang C formula comes from the  $M/M/c/K$  model; for the loss system you will get the Erlang B formula

(Refer Slide Time: 37:49)



The fifth model that is the M/M/infinity; it is not a queuing model, because the servers are infinite, unlimited servers in the system. Therefore, the customer whoever enter he will get immediately the service. The service will be started immediately, whereas, the service time is exponentially distributed with the parameter  $\mu$  by the each server. All the servers are identical, the number of server are infinite here.

Therefore you will have the underling stochastic process for the system size that is the birth-death process with the birth rates are  $\lambda$ , because the population is from the infinite source. The death rates are  $1\mu$ ,  $2\mu$  and so on, because the number of servers are infinite. So, the model which I have discussed in today's lecture all the five models are the underlining stochastic process is a birth-death process; this is the simplest Markovian queuing models.


(Refer Slide Time: 38:59)

**Steady-state Distribution**

$$\pi_n = \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n \pi_0, \quad n=1,2,\dots$$

Using  $\sum_{i=0}^{\infty} \pi_i = 1$ ,  $\pi_0 = e^{-\frac{\lambda}{\mu}}$

Hence,

$$\pi_n = \frac{e^{-\frac{\lambda}{\mu}} \left(\frac{\lambda}{\mu}\right)^n}{n!}, \quad n=0,1,2,\dots$$
$$N \sim \text{Poisson}\left(\frac{\lambda}{\mu}\right)$$



You can get the steady-state distribution; use the same theory of birth-death process, and if you observe this steady-state probabilities is of the same Poisson. It is of the form that is the probability mass function of the Poisson distribution. Therefore, you can conclude in a steady-state the number of customers in the system that is Poisson distributed with the parameter lambda by mu, because the probability mass function for the  $\pi_n$  is the same as the probability mass function of exponential distributed random variable with the parameter lambda by mu.

(Refer Slide Time: 39:44)

**Transient Solution of Finite BDPs**

Transient solution of M/M/1/N, M/M/c/K and M/M/c/c

- Polynomial method (Murphy and O'Donohoe (1975))
- Polynomial method (Rosenlund (1978))
- Matrix method (Chiang (1980))
- Spectral representation method (Van Doorn (1981))
- Orthogonal polynomial method (Nikiforov et al (1991))
- Eigenvalues method (Kijima (1997))



Now I am explaining the transient solution of a finite birth-death process. So, using these one can find out the transient solution of the birth-death process which I have discussed in today's class M/M/1/n, M/M/c/k and M/M/c/c also. So, the logic is same; that means you have a birth-death process with the finite state space. Therefore, the Q matrix is going to be a degree, whatever be the number of states in the state space, and it is going to be a dry diagonal matrix. And you know the lambda n's and mu birth rates as well as the death rates, and the birth rates and death rates are going to be different for these three models.

There are many literatures over the transient solution of a finite birth-death process started with a Murphy and O'Donohoe. He uses the polynomial method and in 1978 Rosenlund also found the transient solution for a finite BDP using again the different polynomial methods. And Chiang in 1980 he made a matrix method to get this transient solution then later Van Doorn gave the solution using a spectral representation method. And Nikiforov et al, 1991, he also gave the transient solution using the orthogonal polynomial, and later Kijima also gave the solution using Eigen value methods. So, these are all the literatures for getting the transient solution of a finite birth-death process.

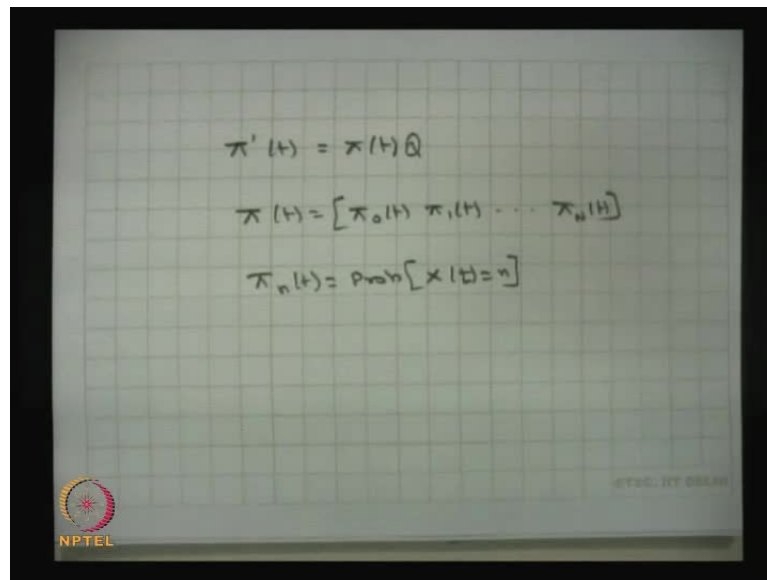
(Refer Slide Time: 41:39)

Transient behaviour of an M/M/1/N Queue  
 - O.P. Sharma and U.C. Gupta  
 Appears in sbch. proc. & their Appl. 13 (1982) 327-331  
 Let  $\psi(n, \theta) = \int_0^{\infty} e^{-\theta t} \pi_n(t) dt$  ;  $\pi_0(0) = 1$   
 $(\lambda + \theta) \psi(0, \theta) = \mu \psi(1, \theta) + 1$   
 $(\lambda + \mu + \theta) \psi(n, \theta) = \mu \psi(n+1, \theta) + \lambda \psi(n-1, \theta)$  ;  $1 \leq n \leq N-1$   
 $(\mu + \theta) \psi(N, \theta) = \lambda \psi(N-1, \theta)$   
 The solution is  
 $\psi(n, \theta) = A \alpha^n + B \beta^n$  ;  $\alpha, \beta = \frac{\theta + \lambda + \mu \pm \sqrt{(\theta + \lambda + \mu)^2 - 4 \lambda \mu}}{2 \mu}$

And here I am going to explain how to get the transient behavior of M/M/1/N queue in a very simplest form even though there are this many literature and many more literatures for the finite birth-death process. But here I am explaining the overview of how to get the

transient behavior of M/M/1/N queue, and this is by O P Sharma and U C Gupta it appears in the stochastic processes and their applications Vol. 13 1982. So, what this method work you start with the forward Kolmogorov equation that is  $\pi'(t) = \pi(t)Q$  and that is started with.

(Refer Slide Time: 42:33)



$$\pi'(t) = \pi(t)Q$$

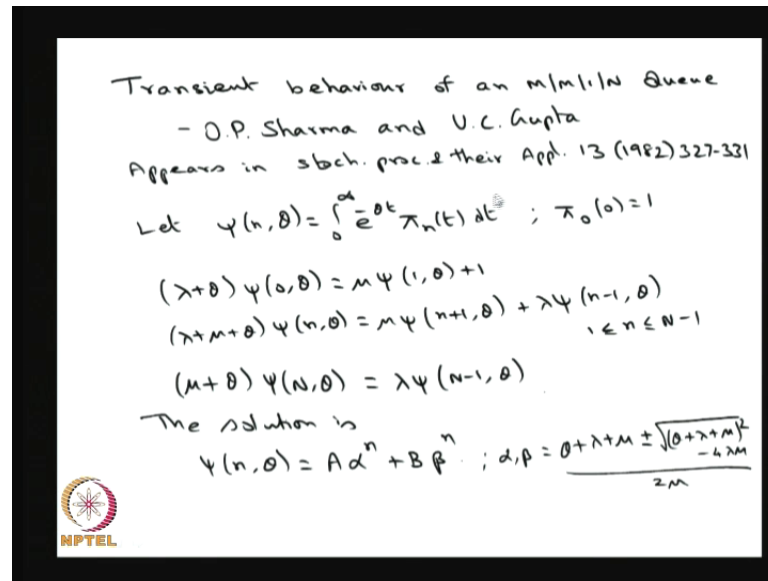
$$\pi(t) = [\pi_0(t) \ \pi_1(t) \ \dots \ \pi_N(t)]$$

$$\pi_n(t) = \text{Prob}[X(t)=n]$$

$\pi'(t)$  that is equal to  $\pi(t)$  into  $Q$  matrix where  $\pi$  is the matrix, and  $\pi'$  is with the derivatives and  $Q$  is the infinite decimal matrix. Take a forward Kolmogorov equation then use the Laplace transform and so on for each  $\pi_n(t)$  you take, sorry here the  $\pi'(t)$  is the vector it is a distribution of a  $X$  of  $t$ . Therefore, this is a vector, and this is the vector, and  $Q$  is the matrix not the matrix which I said wrongly. So, this is a vector, and this is a vector, and  $Q$  is the matrix.

So, take a Laplace transform for each probability where the  $\pi_n(t)$  that is nothing but so the  $\pi(t)$  is the vector that started with  $\pi_0(0)$   $\pi_1(0)$  and so on  $\pi_n(0)$  where  $\pi_n(t)$  is nothing but what is the probability that, the same notation I started when I discussed the continuous time Markov chain, what is the probability that  $n$  customers in the system at time  $t$ . These are unconditional probability distribution. So,  $\pi_n(t)$  is the probability that the  $n$  customers in the system are time  $t$  and using  $\pi_n(t)$  you get the vector, and you make a forward Kolmogorov equation  $\pi'(t)$  is equal to  $\pi(t)$  times  $Q$ .

(Refer Slide Time: 44:15)



Transient behaviour of an M/M/1/N Queue  
 - O.P. Sharma and U.C. Gupta  
 Appears in sbch. proc. & their Appl. 13 (1982) 327-331


Let  $\psi(n, \theta) = \int_0^\infty e^{-\theta t} \pi_n(t) dt$  ;  $\pi_0(0) = 1$

$$(\lambda + \theta) \psi(0, \theta) = \mu \psi(1, \theta) + 1$$

$$(\lambda + \mu + \theta) \psi(n, \theta) = \mu \psi(n+1, \theta) + \lambda \psi(n-1, \theta) \quad 1 \leq n \leq N-1$$

$$(\mu + \theta) \psi(N, \theta) = \lambda \psi(N-1, \theta)$$

The solution is

$$\psi(n, \theta) = A \alpha^n + B \beta^n ; \alpha, \beta = \frac{\theta + \lambda + \mu \pm \sqrt{(\theta + \lambda + \mu)^2 - 4\lambda\mu}}{2\mu}$$


And take a Laplace transform for each  $\pi_n$  of  $t$  that exist, because this is the probability and the conditions for the Laplace transform of these functions satisfies, you can cross check; therefore, you are taking a Laplace transform and this is going to be a function of theta. Before taking a Laplace transform you need an initial condition also. So, at time zero you assume that there are no customer in the system, at time zero no customer in the system; that means  $\pi_0$  of 0 is equal to 1. Therefore, that probability is going to be 1, and all other probabilities are going to be zero; that is the initial probability vector. So, use this initial probability vector and apply it over the forward Kolmogorov equation, taking a Laplace transform you will get the system of algebraic equation.

Since you are using a  $\pi_0$  of 0 is equal to 1 you will get the first equation with the term 1 and all other terms are going to be 0, and you know the Laplace transform of the derivative of the function. So, you substitute you take a Laplace transform over the forward Kolmogorov equation with this initial condition as well as  $\pi_n$ 's of 0 is equal to 0 for  $n$  not equal to 0. So, you will have an algebraic equation that is  $n$  plus 1 algebraic equation is a function of theta.

You have to solve this algebraic equation system of algebraic equations in terms of theta. Once you are able to solve this and take an inverse Laplace transform and that is going to be the system size at any time  $t$ . You can start saying that this is going to be of the solution of  $A$  times  $\alpha^n$  and  $B$  times  $\beta^n$  where  $\alpha$  and  $\beta$  are given in

this form where alpha is equal to this plus something and the beta is equal to minus something minus square root of this expression. So you will have alpha as well as beta. Now what you want to find out? If you find out the constant A and B you can get the Laplace transform of  $p_n$  of  $t$  then you take an inverse Laplace transform and you get the  $p_n$  of  $t$ .

(Refer Slide Time: 46:41)

$$D(\theta) = \begin{vmatrix} \theta + \lambda & \mu & & & \\ \lambda & \theta + \lambda + \mu & \mu & & \\ 0 & \lambda & \theta + \lambda + \mu & \mu & \\ & & \ddots & \ddots & \ddots \\ & & & \lambda & \theta + \lambda + \mu & \mu \\ 0 & & & & \lambda & \theta + \lambda & \mu \end{vmatrix}_{n+1}$$

$$= \theta \varphi_n(\theta)$$

where  $\varphi_n(\theta) = \prod_{k=1}^n (\theta + \lambda + \mu + d_{n,k} \sqrt{\lambda \mu})$

$d_{n,k}$  -  $k^{\text{th}}$  roots of  $n^{\text{th}}$  degree Chebyshev's polynomial of second kind  $U_n(x)$ .

Note that  $\varphi_n(\theta)$  has distinct real factors.

So, for that you need the determinant of matrix of this form, and here this is nothing but all these values are death rates, and these are all the birth rates, and this is corresponding to the M/M/1/N model, and the same logic goes for the transient solution of the M/M/c/K as well as M/M/c/c. So, instead of this lambdas and mu's you will have a corresponding birth rates and the death rates, but ultimately you will have a  $n$  plus 1 matrix determinant as a function of  $\theta$ . And since these three models are going to be a irreducible positive recurrent the stationary probability and the limiting probabilities exist; therefore, this determinant is going to be always of the form  $\theta$  times some other function as a degree as a polynomial of degree  $n$  in a function of  $\theta$ .

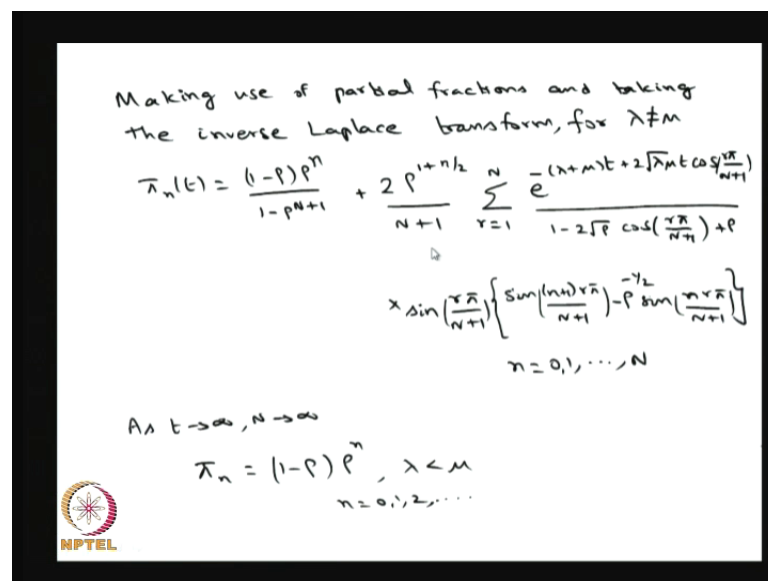
So, this  $\theta$  is corresponding to the stationary probabilities or the limiting probabilities. Therefore, always you can get the  $n$  plus 1 th degree matrix 1 th order matrix determinant that is  $\theta$  times the polynomial of degree  $n$  as a function of  $\theta$ . For the M/M/1/N model the birth rates are  $\lambda$  and the death rates are  $\mu$ , and you can get this polynomial also in the form of product.



The product of theta plus lambda plus mu times alpha of N comma k square root of lambda mu where alpha of N comma k is nothing but the k roots of a n th degree Chebyshev's polynomial of second kind. There is the relation between the birth-death processes with the orthogonal polynomial. For instance the M/M/1/N model the finite capacity M/M/1/N model the corresponding orthogonal polynomial for this birth-death process is the Chebyshev's polynomial of the second kind.

Similarly you can say the orthogonal polynomial corresponding to the M/M/c/c model that is the Charlier polynomial. Like that we can discuss the corresponding orthogonal polynomial for the finite capacity birth-death processes. So, here for the M/M/1/N model this is related to the Chebyshev's polynomial of second kind that is  $U_n$  of  $x$ . So, once you are able to get the Chebyshev's polynomial roots and that root is going to play a role in the product form and that is going to be polynomial. Note that this polynomial has a distinct real factor.

(Refer Slide Time: 49:55)



Making use of partial fractions and taking the inverse Laplace transform, for  $\lambda \neq \mu$

$$P_n(t) = \frac{(1-\rho)\rho^n}{1-\rho^{N+1}} + \frac{2\rho^{1+n/2}}{N+1} \sum_{r=1}^N \frac{e^{-(\lambda+\mu)t + 2\sqrt{\lambda\mu}t \cos(\frac{r\pi}{N+1})}}{1 - 2\sqrt{\rho} \cos(\frac{r\pi}{N+1}) + \rho}$$

$$\times \sin\left(\frac{r\pi}{N+1}\right) \left\{ \sin\left(\frac{(n+1)r\pi}{N+1}\right) - \rho \sin\left(\frac{nr\pi}{N+1}\right) \right\}$$

$$n = 0, 1, \dots, N$$

As  $t \rightarrow \infty, N \rightarrow \infty$

$$P_n = (1-\rho)\rho^n, \quad \lambda < \mu$$

$$n = 0, 1, 2, \dots$$

NPTEL

Therefore, you can use the partial fraction, then you take an inverse Laplace transform and finally you can get the  $p_i$  of  $t$ . I am skipping all the simplification part and the main logic is this  $n$  plus 1 th order matrix determinant and that determinant has the factors and those factors are related to the Chebyshev's polynomial roots. So, once you use all those logics and use the partial fraction then finally you take an inverse Laplace transform.

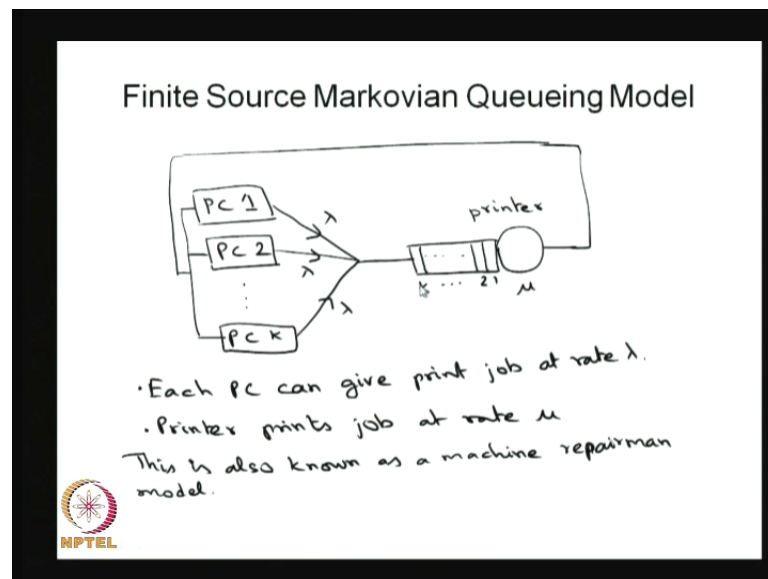
For  $\lambda$  is not equal to  $\mu$  you will get a steady-state or stationary probabilities plus this expression, and this is the function of  $t$   $e^{\mu t}$  minus  $\lambda$  plus  $\mu$  times  $t$  plus  $2$  times square root of  $\lambda \mu$  times  $t \cos$  of  $r \pi$  by  $n$  plus  $1$  and denominator this expression multiplied by this, and here this result is related to the initial condition  $0$ ; that means a time  $0$  the system is empty. If the system is not empty then you will have one more expression here  $\sin$  of this minus another term.

So, that is you will have a getting bigger expression for system size is not empty, and this  $\theta$  times this that will give the corresponding partial fraction and so on, inverse Laplace it will give the terms which is independent of  $t$  and that is the related to the steady-state probabilities, because if you put  $t$  tends to infinity if the quantities are greater than  $0$ , so as  $t$  tends to infinity the whole terms will tend to zero. Therefore, as  $t$  tends to infinity you will have  $P_n$  of  $t$  is equal to this expression, and this is valid for  $\rho$  is less than  $1$ ; with that condition  $\rho$  is less than  $1$  those terms will tends to  $0$ .

And you will have only this term and that is going to be the steady-state or limiting probabilities for  $M/M/1/N$  model. If you make also  $n$  tends to infinity along with the  $t$  tends to the infinity you will have  $P_n$ 's that is the steady-state probability for the  $M/M/1$  infinity model. So, even though I have explained the  $M/M/1/N$  transient solution in a brief way but the same logic goes for the  $M/M/c/c$  model also. The only difference is this determinant has the  $\lambda$ 's and instead of  $\mu$ 's you will have  $\mu^2$ ,  $\mu^3$ ,  $\mu$  and so on.

And instead of the Chebyshev's polynomial you will land up with the Charlier polynomial, but there is a difference between this  $M/M/1/N$  model and the  $M/M/c/c$  model transient solution. Since the Chebyshev's polynomial has a closed form roots you can find out the factors. So, here these are all the factors, and you know the factors as well as you can get the closed form expression for the  $M/M/1/N$  transient solution whereas the Charlier polynomial does not have a closed form roots. Therefore you will land up with the numerical result for the transient solution for  $M/M/c/c$  model.

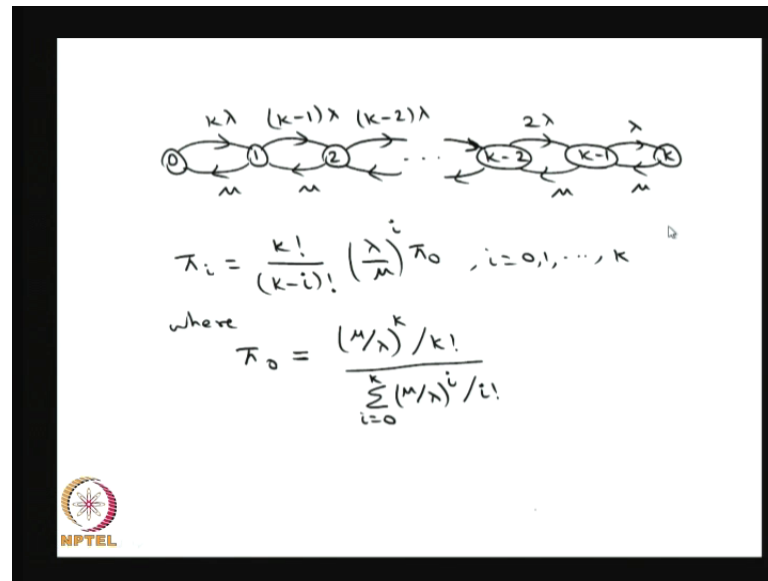
(Refer Slide Time: 53:31)



Location of a continuous time Markov chain that is a finite source Markovian queueing models. This model is also known as a machine repairman model, and you can think of this PCs are nothing but the machines, and this is nothing but the repairman, and here the scenario is we have KPCs and each PC can give a print job and the inter arrival of print jobs that is exponentially distributed by the each PC. Therefore the print jobs that follow a arrival process that is the Poisson process with the parameter  $\lambda$  from each PC. And once the print jobs comes into the printer it will wait for the print and the time taken for the each print that is also exponentially distributed with the parameter  $\mu$ , and here there is an another assumption before the first print is over by the same PC it cannot give another print command.

Therefore, after the print is over by any one particular print job of any PC then these things will go back to the same thing. Then with the inter arrival of print jobs generated that is exponentially distributed then the print job can come into the printer. So, with these assumptions you can think of the stochastic process; that means number of print jobs at any time  $t$  in the printer that is going to form a stochastic process, and with the assumption of inter arrival of print jobs that is exponential and the actual printing job that is exponentially distributed and so on.

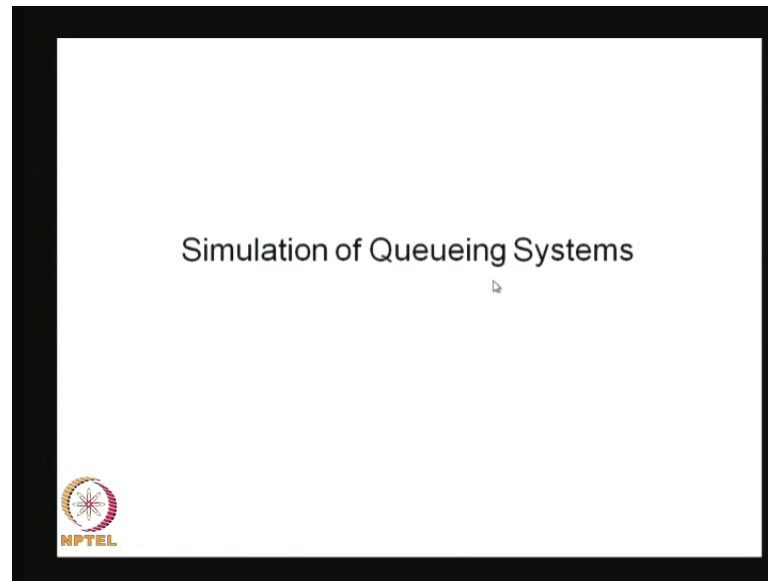
(Refer Slide Time: 55:34)



Therefore this is going to be a birth-death process with the birth rates or  $k$  times  $\lambda$  and  $k$  minus 1 times  $\lambda$  and so on, whereas, the death rates that is  $\mu$  because we only one repair. So, this is nothing but system size number of jobs in the print job printer, so therefore, that varies from 0 to capital  $k$  because we are making the assumption more than one print job cannot be given by the same PC before the print is over, and from 0 to 1 the arrival rate will be any one of the KPCs; therefore, the arrival rate is  $k$  times  $\lambda$  and already one print job is there in the system printer. Therefore, out of  $k$  minus 1 PC is one print job can come; therefore, the inter arrival time that is exponentially distributed with the parameter  $k$  minus 1 times  $\lambda$  and so on. So, this is the way you can visualize the birth rates whereas the death rates are  $\mu$ .

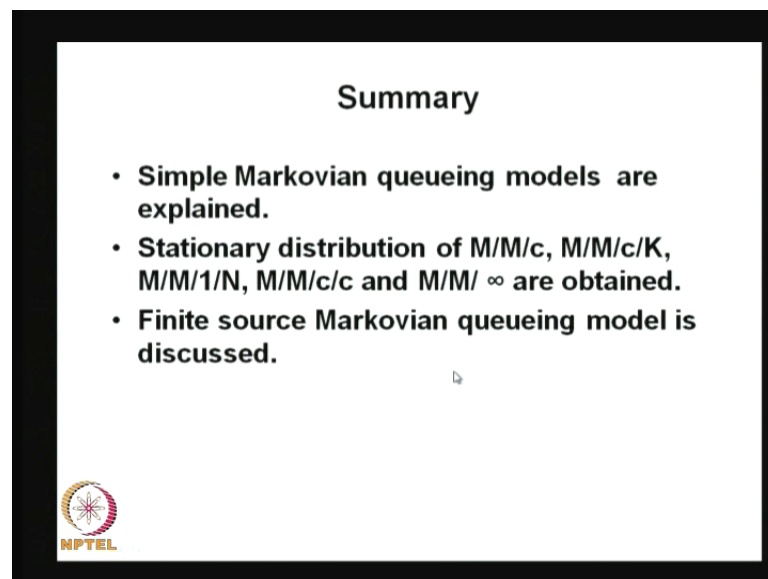
Once you know the birth rates and the death rates you can apply the birth-death process concept to get the steady-state probabilities. So, here we are getting the  $\pi_i$ 's in terms of  $\pi_0$  and using the summation of  $\pi_i$  is equal to 1 you are getting the  $\pi_0$  also. And once you know the steady-state probability you can get all other measures. So, the difference is in this model it is finite source; therefore, the birth rates are the function of it is a state-dependent birth rates whereas the death rates are  $\mu$ 's only.

(Refer Slide Time: 57:04)



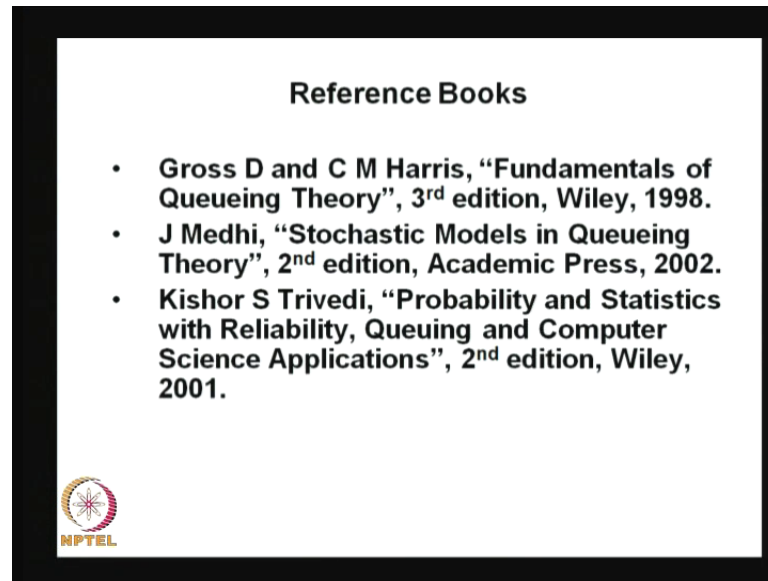
Simulation of a queueing model I will do it in the next lecture.

(Refer Slide Time: 57:10)



The summary of today's lecture I have discussed the simple Markovian queueing models other than  $M/M/1/\infty$  that I have discussed in the previous lecture and stationary distribution and all the other performance measures using the birth-death process we have discussed for this queueing models, and finally I discussed the finite source Markovian queueing model also.

(Refer Slide Time: 57:37)



These are all the reference books.

Thanks.