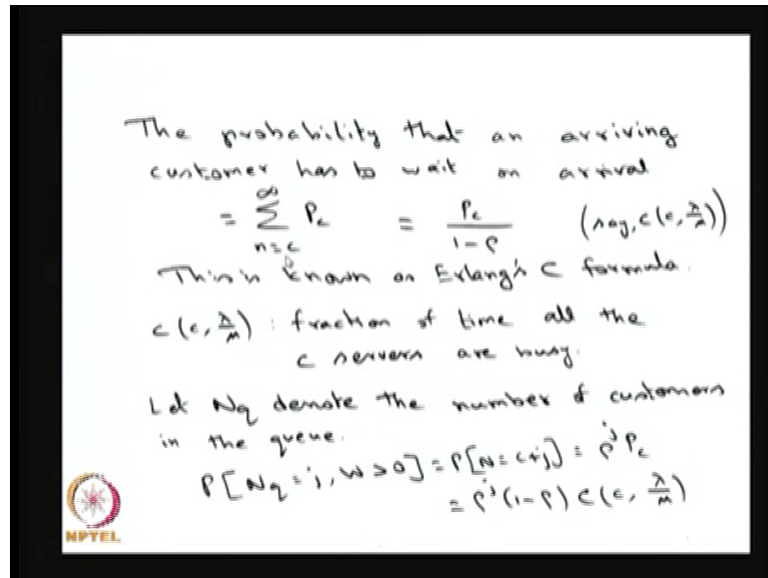


**Introduction to Probability Theory and Stochastic Processes**  
**Prof. S. Dharmaraja**  
**Department of Mathematics**  
**Indian Institute of Technology, Delhi**

**Lecture – 90**

(Refer Slide Time: 00:01)



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 or arrival? So, that means, the number of customers in the system is greater 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 a some it mistake  $P_n$  suffix  $n$ , where  $n$  is running from  $c$  to infinity. If you add all those probabilities, that is going to be  $P_c$  divided by  $1 - \rho$ .

And this probability is known as a Erlang's  $c$  formula. For a multi server, infinite capacity model that I am denoting with the letter  $c$  of  $c$  comma  $\lambda$  by  $\mu$ . Because, you need a number of servers in the system and you need a  $\lambda$  as well as  $\mu$ . If I know this quantity, I can find out what is a Erlang  $c$  formula. This is very important formula.

Using that, you can find out what is the optimal  $c$ , such a way that and the probability has to be minimum. You can find out what is optimal number of servers is needed to

have a some upper bound probability of arriving customer has to wait. Therefore, this Erlang C formula is very useful in performance analysis of any system.

The next quantities  $N_q$  denotes the number of customers in the queue. So, either I use the letter  $N$  suffix  $q$ . Earlier I use 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  $q$  is  $j$  and the waiting time is going to be greater than 0.  $W$  is used for the waiting time. So, the waiting time is going to be greater than 0; 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 0 that is same as what is the probability that  $c$  plus  $j$  customers in the system. Do the little simplification. So, you will get this joint probability in terms of Erlang C formula.

(Refer Slide Time: 02:44)

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  $j$  customers in the queue? Given that, at the waiting time is greater than 0, if we do little simplification, I will get a 1 minus 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.

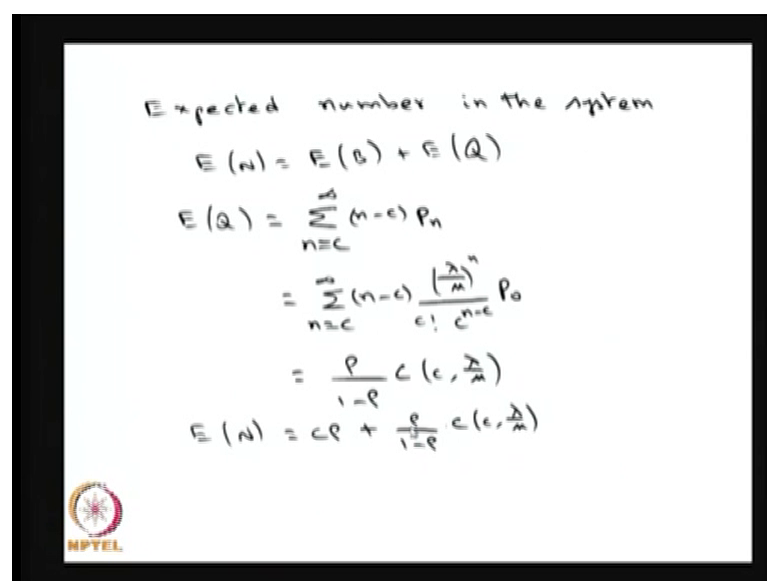
From this, we can find out the expected number of this next measure is expected number of busy service. What is the average number of a busy service? That is nothing, but the

summation of  $n$  equal to 0 to  $c$  minus 1  $n$  times  $P_n$ ; that means, whenever the system size is less than  $c$ , only those many servers are busy. And with the probability, whenever a  $n$  customer or more than  $n$  customers in the system, all the  $c$  servers are going to be busy. Therefore,  $c$  times  $P$  you simplify you will get a  $c$  times  $\rho$ . So, that is the expected number of busy service.

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 a expected number of idle server is nothing, but expectation of a 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  $B$ . Therefore,  $c$  minus  $B$  is same as  $I$ .

So, the expectation 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  $c$  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  $1$  minus  $\rho$ . So, other than stationary distribution for the MMC 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 0 as well as this expected quantities we are getting.

(Refer Slide Time: 05:24)



Expected number in the system


$$E(N) = E(B) + E(Q)$$

$$E(Q) = \sum_{n=c}^{\infty} (n-c) P_n$$

$$= \sum_{n=c}^{\infty} (n-c) \frac{(\frac{\lambda}{\mu})^n}{c! c^{n-c}} P_0$$

$$= \frac{\rho}{1-\rho} c \left( c, \frac{\lambda}{\mu} \right)$$

$$E(N) = c\rho + \frac{\rho}{1-\rho} c \left( c, \frac{\lambda}{\mu} \right)$$



Also, we can find out what is a expected number of customers in the system. That is nothing, but expected number is nothing, but expected of the busy servers plus a expected number in the Q. Earlier, I use the notation n suffix Q n suffix Q and Q are both one and the same.

So, I can compute what is the expectation of Q. Do the little simplification. Then I can substitute expectation of Q here. Therefore, I will get expected number of customers in the system that involves the Erlang C formula. So, this Erlang 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.

(Refer Slide Time: 06:27)

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

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

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

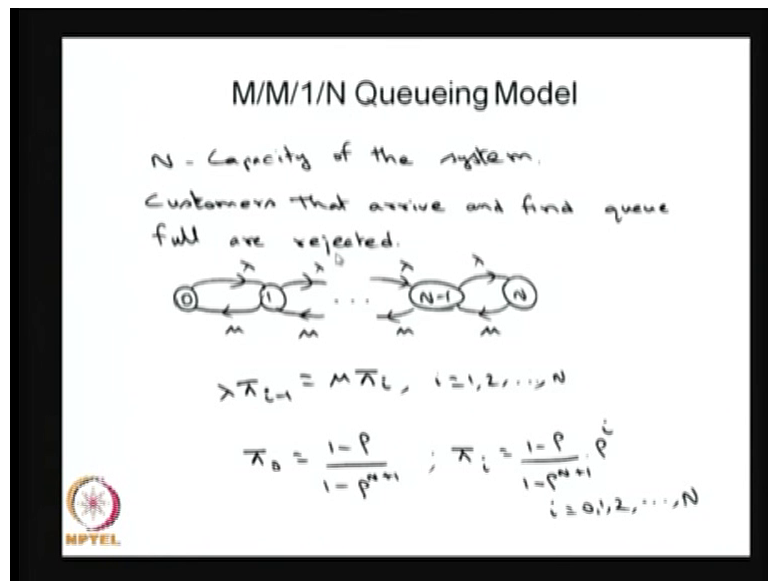
So, using a Little's formula, using Little's formula, I can find out the expected time spent in the system. Because, I know what is arrival rate and from the stationary distribution, I got expected number in the system in a steady state. Therefore, since I know lambda and expectation of a n, I can get expectation of R where R is the response time or sojourn time or 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. You can apply the Little's formula in the q level also. So, this is a system level and you can apply the Q level also. Say lambda times

expectation of a waiting time is same as expectation of a number of customers in the queue.

So, expectation of a waiting time or average waiting time is same as expectation of a Q divided by lambda. So, using since the mm MMC infinity Q the underlying stochastic process is a birth death process. Therefore, we are getting all the measures using the birth death logic.

(Refer Slide Time: 07:55)



Next I am going for the finite capacity.

So, the  $n$  is a capacity of the system ; that means, whenever the customers arrives and find a 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 0 to  $N$ . Therefore, the state space is finite the number of customers in the system may 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 exponentially distributed, only one server finite capacity. Therefore, the underlying stochastic process is a birth death process with the birth rates lambda the death rates mu. If you see the Q matrix, for this one infinitesimal generator matrix that is a tri diagonal matrix with the all the off diagonals or lambdas as well as mu and the diagonals are minus lambda plus mu

except the first term and the last term. The except the first row and the last row, our interest is to get the stationary distribution. Later, I am going to explain the time dependent solution also.

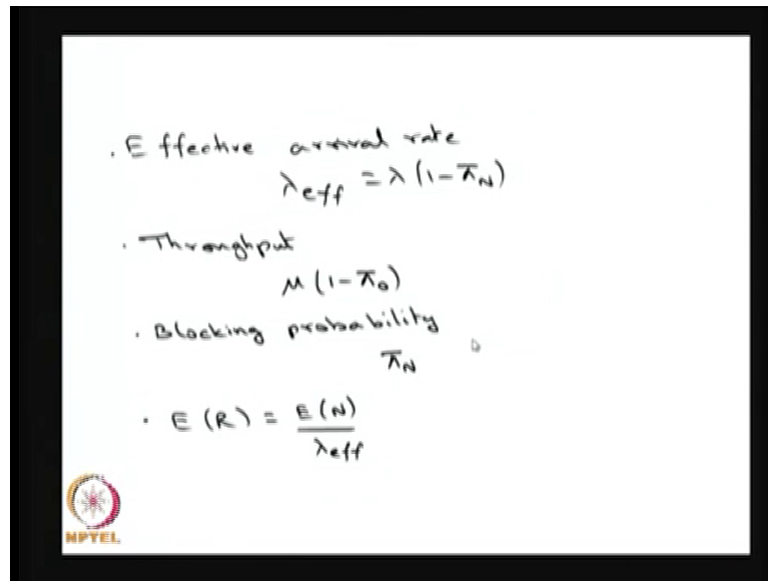
So, to get the stationary distribution, either you write the  $\pi Q$  is equal to 0 and the summation of  $\pi$  is equal to 1 and solve that or you write the balance equation the  $\pi Q$  is equal to 0. That will land up a balance equation. So, some books writes this as a balance equation. What is the inflow rate and what is the outflow rate? Both are going to be same whenever the system reaches the equilibrium solution equilibrium state.

Therefore, the outflow is a  $\lambda$  times, this the inflow is  $\mu$  times  $\lambda$ . Like that you can go for understanding the balance equation for this state and second and so on. And this also satisfies the time. This is also called a satisfying the time reversible equation. Therefore, one can use the time reversible property of a birth death process. So, we can find out the  $\pi$  is easily using the time reversible equation itself. You do not want to use the  $\pi Q$  is equal to 0. Instead of that, you can write the time reversible equation since it is has fed by all the states.

Now, we can use the summation of  $\pi$  is equal to 1,  $i$  starting from 0 to  $n$ . Therefore, you will get  $\pi_0$  and here the birth death process with the finite state space. Therefore, the  $\pi_0$  will be 1 divided by the denominator series that is a finite series finite terms in it. Therefore, it is always converges immaterial of the value of  $\lambda$  and  $\mu$ . Therefore, you will get a  $\pi_0$  without any restriction over  $\lambda$  and  $\mu$ .

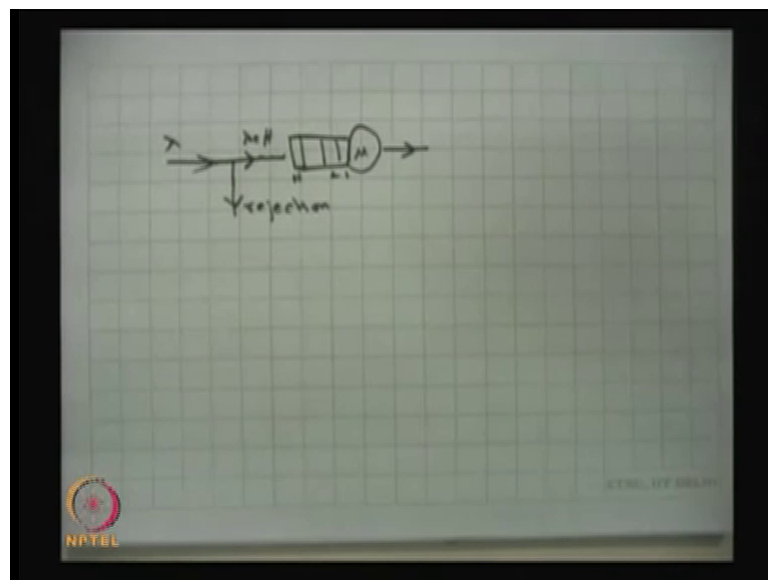
So, once you get the  $\pi_0$ , you can get  $\pi_i$ 's in terms of  $\pi_0$ . Therefore, that is a  $1 - \rho$  divided by  $1 - \rho^{n+1}$  times  $\rho^i$  where  $\rho$  is  $\lambda$  by  $\mu$ . So, this is the birth the underlying stochastic process is the birth death process with the birth rates  $\lambda$  and the death rates  $\mu$ . So, you can use all the concepts of the birth death process and you can analyze the system in a ECV. So, this is a steady state probability.

(Refer Slide Time: 11:57)



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 finite capacity.

(Refer Slide Time: 12:15)



So, maximum  $n$  customers can wait in the 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 nothing, but what is the rate in which the system is the customers are entering into the system. So, there is a partition here.

So, the effective arrival rate is a  $\lambda_e$  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  $\lambda_e$  effective. Whenever the system is not full, that proportion of the time or the probability is  $1 - p_N$  where  $p_N$  is the steady state probability. Just now, we got it from here you can get a  $p_N$  suffix  $N$  that is the probability that the system is full and  $1 - p_N$  is the probability that the system is not full and multiplied by the arrival rate that is going to be the  $\lambda_e$  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  $1 - p_0$ . Therefore,  $1 - p_0$  times  $\mu$  that is the rate in which the customers are served in the  $M/M/1$  system. Whenever the system is not full, not empty, sorry 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  $1 - p_0$ , you can get in terms of  $\lambda$ s equivalent also. But, the 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 is same as the probability that the system is full, that is  $p_N$ . Once we 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 custom divided by not  $\lambda$ . It is  $\lambda_e$  effective.

Because the effective arrival rate is used in the Little's formula, not the arrival rate for a  $M/M/1$  infinity system, the effective arrival rate and the arrival rate are one and the same because, there is no blocking. Therefore, the probability of  $1 - p_N$  that is equal to 1 only. Therefore, the effective arrival rate and the arrival rate are same for a 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 mm lambda effective for the MMC k model also. So, other than stationary distribution or equilibrium probabilities, we are getting the other performance measures using the birth death process concepts.