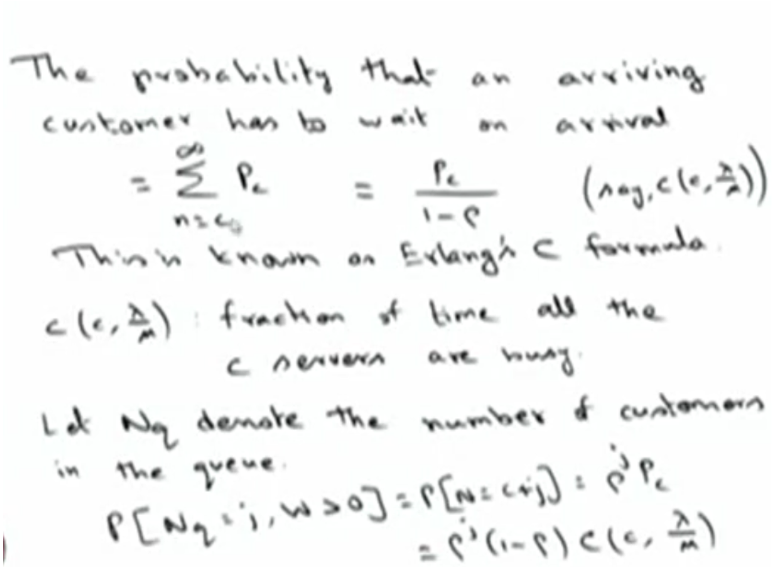
Stochastic Processes - 1
Dr. S. Dharmaraja
Department of Mathematics
Indian Institute of Technology – Delhi

Lecture - 66
M/M/1/N Queueing Model

(Refer Slide Time: 00:00)



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 on 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 suffix n where n is running from c to infinity.

If you had all those probabilities, that is going to be Pc divided by one minus rho. This probability is known as a Erlang C formula for a multi server infinite capacity model. That I am denoting with the letter c comma lambda by u, because you need number of service in the system and you need lambda as well as mu. If I know this quantity, I can find out what is Erlang's C formula. This is 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 service is needed to have 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 quantity is n q denotes the number of customers in the queue. So either 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 customers in the system. Do the simplification, so you will get this joint probability in terms of Erlang's C formula.

(Refer Slide Time: 02:45)

Thus,

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

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

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. What is the conditional probability that j customers in the queue given that the waiting time is greater than zero. If you do little simplification, I will get one minus rho times rho power j where rho is lambda divided by c mu. This is nothing but the probability mass function of geometric distribution.

This is the probability mass function of a geometric distribution; therefore, this conditional probability is geometrically distributed with a parameter rho. From these we can find out the expected number of, 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 minus one 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 customer are 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 c times rho, 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, (()) (04:20) that is expected number of idle server is nothing but expectation of, it is a random variable. So ideal 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 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 B expectation of B just now we got c times rho, therefore the expected number of idle server is c times one 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 zero as well as this expected quantities we are getting.

(Refer Slide Time: 05:25)

Expected number in the system

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

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

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

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

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

Also we can find out what is 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 and q are both one the same. So I can compute what is the expectation of q, it is a little simplification and then I can substitute expectation of q here.

Therefore, I will get expected number of customers in this 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 later you can do some optimization over the probability expected number with specified C and lambda by mu.

(Refer Slide Time: 06:26)

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

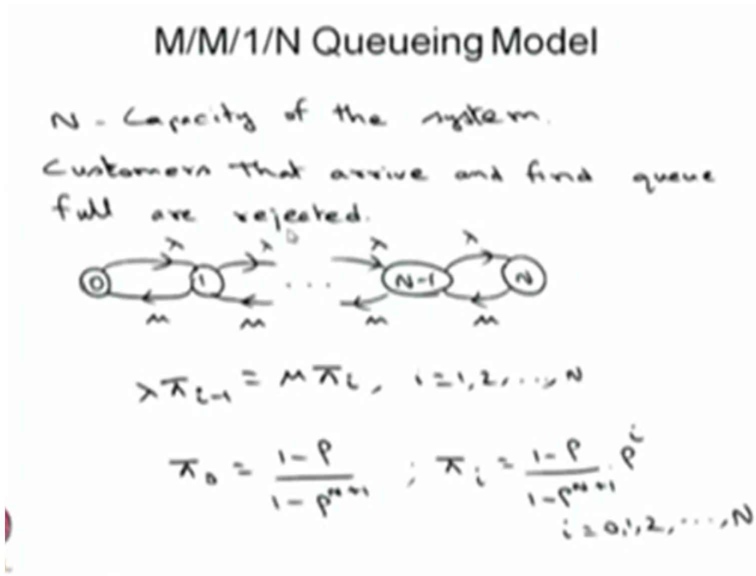
$$E(W) = \frac{E(Q)}{\lambda} = \frac{P_c}{c\mu(1-\rho)^2}$$

So using little's formula I can find out the expected time spend in the system because I know what is the arrival rate and from the stationary distribution I got expected number in the system in the steady state. Therefore, since I know lambda and expectation of n, I can get expectation of r, where r is the response time or sojourn time or total time spend in the system.

So that expectation is going to be, expectation of n divided by lambda. Do a 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. So lambda times expectation of waiting time is same as expectation of number of customers in the queue.

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

(Refer Slide Time: 07:56)



Next I am going for the finite capacity. So the n is the capacity of the system, that means when the customers arrives and find q 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, that random variable takes the possible values from zero to capital N. So the states phase is fine. The number of customers in the system is anytime t.

That is a random variable and you will have a stochastic process. And since the entire arrival time is exponentially distributed services, exponentially distributed only one server, finite capacity, therefore the underlying stochastic process is birth-death process, if the birth rate is lambda and the death rate is mu. If you see the q matrix for case 1, infinitesimal generated matrix.

That is a dry diagonal matrix with all the off diagonals or lambdas as well as mu and diagonals are minus lambda plus mu except the first term and last term. Except the first row and last row. Our interest is to get the stationary distribution later I am going to explain the time dependence relation also. So to get the stationary distribution either you write q equal to zero.

And the summation of pi equal to one and solve that or you write the balance equation the pi q equals to zero 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 equilibrium state. Therefore, the outflow is lambda times this and inflow is mu times lambda one, like that you can go for understanding the balance equation for the state and second and so on.

And this also satisfies the, this is also called satisfying the time reversible equation. Therefore, one can use the time reversible property of a birth-death process. So you can find out the pi is easy using the time reversible equation itself. You do not want to use 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 one, i starting from zero on n, therefore you will get pi not and here the birth-death process with the finite state space, therefore the pi not will be one divided by the denominator series, that is the finite series, finite terms in it. Therefore, it is always converges immaterial of the value of lambda and mu. Therefore, you will get pi not without any restriction over lambda and mu.

So once you get the pi not, you can get pi i in terms of pi not therefore that is one minus rho divided by one minus rho power n plus one times rho power i where rho is lambda by mu. So this is the underlying stochastic process as birth-death process with the birth rates lambda and death rate is 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: 11:57)

Effective arrival rate

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

- Throughput

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

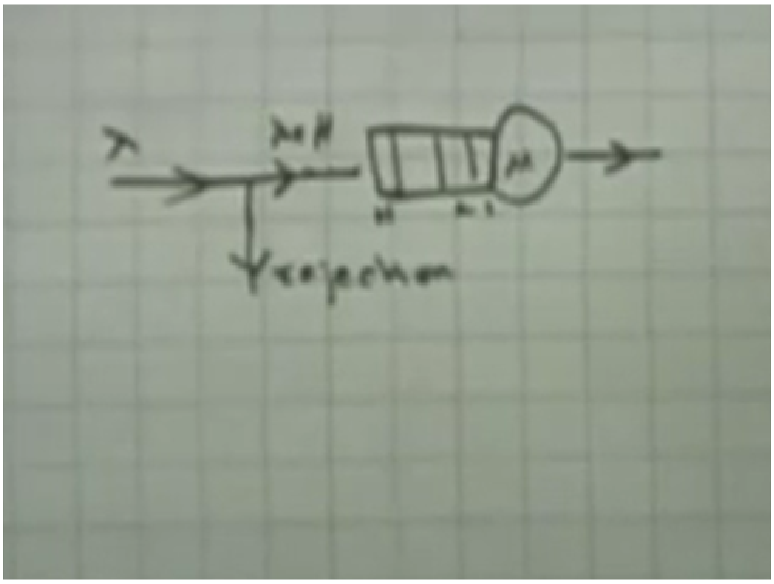
- Blocking probability

$$\pi_N$$

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

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

(Refer Slide Time: 12:15)



So maximum n customers can weight 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 customer is rejected, there for 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 system is, the customers are entering into the system.

So there is a partition here. So the effective arrival rate is lambda. 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 effect. Whenever the system is not full, that proportion of the time or the probability is one minus pi n where pi n is the steady state probability, just now we got it.

From here you can get pi suffix n that is probability that the system is full, that is one minus pi n is the probability is that the system is not full and multiplied by the arrival rate that is going to be the lambda effect. And you can also find out the throughput. Throughput is nothing but, what is the rate in which the customers are served per any tough time. The service rate is mu and this is the probability that the system is not empty, one minus pi not.

Therefore, one minus pi not times mu that is the rate in which the customers are served in the MM1N system. Whenever the system is not empty, that probability multiplied by mu that is going to be the throughput. By using the time reversible equation, the mu times one minus pi not you can get in terms of lambda 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 pi.

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 spend 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 a MM1 infinity system, the effective arrival rate and the arrival rate are one and the same because there is no blocking, therefore the probability of one minus pi n that is equal to one only.

Therefore, the effective arrival rate and the arrival rate are same for 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 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.