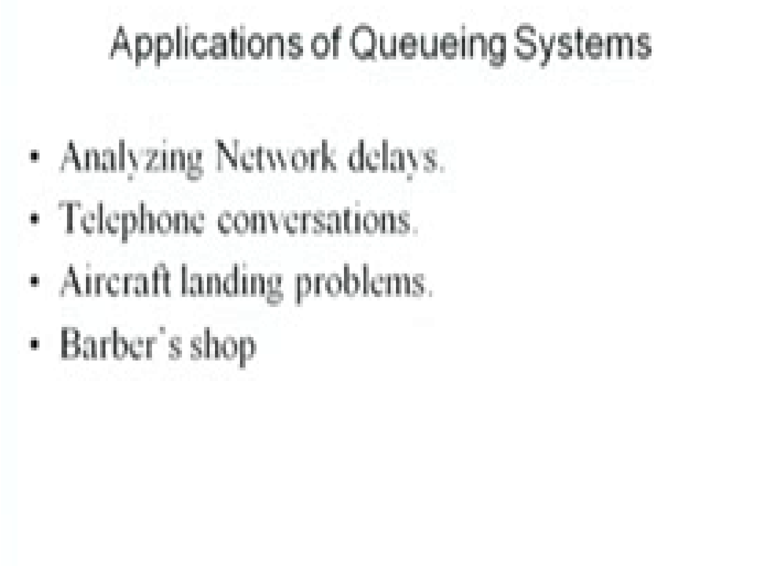
Stochastic Processes - 1
Dr. S. Dharmaraja
Department of Mathematics
Indian Institute of Technology - Delhi

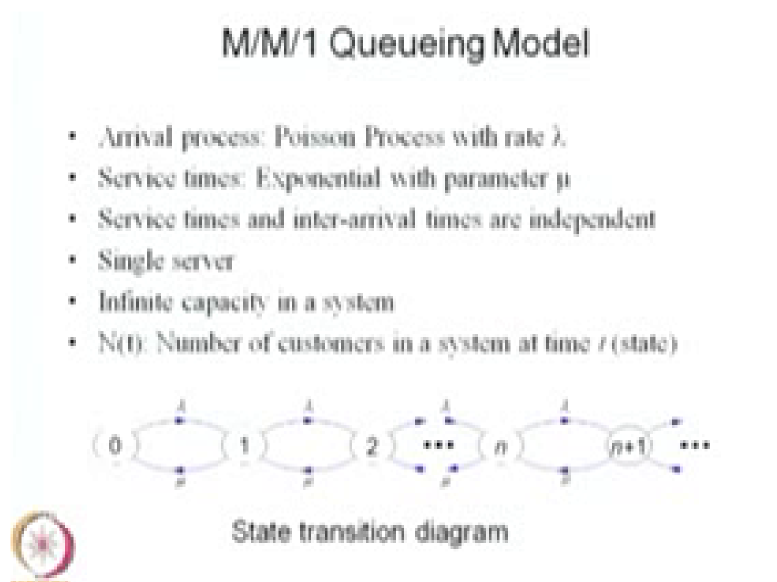
Lecture - 62
M/M/1 Queueing Model

(Refer Slide Time: 00:00)



There are many applications of queuing system, we are going to discuss the abstract queuing system in the further lecture.

(Refer Slide Time: 00:13)



The easiest or the simplest queuing model that is a Markovian queuing model that is MM1 queuing model, later we are going to correlate to the birth death process also, in the MM1 queuing model the inter arrival time is exponentially distributed as I discussed the Poisson process in the previous lecture whenever you have arrival follows Poisson process then the inter arrival time follows exponential distribution and are independent also.

So here the first information is arrival process follows a Poisson process with intensity or rate lambda, that means the inter arrival times are independent and each one is exponentially distributed with a parameter lambda. The second information that is service time, service times are exponentially distributed with a parameter mu and the service times are independent for each customers and that's also independent with the arrival process.

That means there is no dependency over the arrival pattern with the service pattern, service times and inter arrival times are independent. Then the third information only one server in the system that’s a queuing system in which only one server. And the fourth information is missing that means it is a default its infinite capacity model - infinite capacity model.

Now our interest is to find out the behavior of a queuing system or the behavior of number of customers in the system at any time t therefore you can define a random variable N of t, that is nothing but the number of customers in the system at time t, therefore this is going to follow form a stochastic process over the t, since the inter arrival time is exponentially distributed and the service times are exponentially distributed.

The memory less properties is going to be satisfied throughout all the time, therefore this stochastic process there is a discrete state continuous time stochastic process satisfying the Markov property, therefore this is a Markov process, since inter arrival time is exponentially distributed and the service time is exponentially distributed and both are independent and the service time is also independent for each customers.

Therefore this stochastic process satisfies the memoryless property at all-time points, therefore these discrete state because the possible values of N of t, since it is a number of customers the possible values are 0 1 2 and so on, countably infinite therefore it is a discrete state and you are observing the system over the time therefore it is a continuous time.

Therefore this stochastic process is the discrete state continuous time stochastic process satisfying the Markov property based on these assumptions, therefore N of t is a Markov process since the state space is a discrete therefore this is a Markov chain, therefore this is a continuous time Markov chain, therefore N of t is a CTMC, so one can write the state transition diagram for the - for this CTMC.

That means the possible states are 0 1 2 and so on, so this will form a nodes and you try to find out what is the rate in which the system is moving from one state to other state, since it is a MM1 queue model queuing model, therefore whenever the system is in the - whenever the system is in the state 0 by the inter arrival time which is exponentially distributed the number of customers in the system will be incremented by 1.

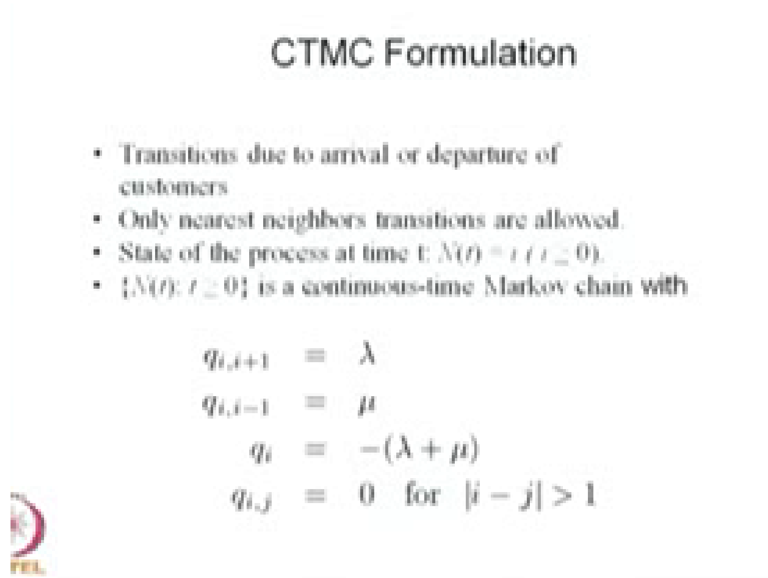
Therefore that rate will be lambda or the system moving from the state 0 to 1 it spends exponentially distributed amount of time here before moving into the state 1, once the system come to the state 1 either one more arrival is possible or the customer who is under service then service could have been finished, therefore the service time is exponentially distributed with a parameter mu, therefore the system goes from the state 1 to 0 with a parameter mu.

Similarly from 1 to 2 because of the inter arrival time is exponentially distributed with the parameter lambda, therefore this is lambda, since the arrival follows a Poisson process in a very small interval of time only one customer is possible with the probability lambda times delta t and so on, therefore there is no way the system goes from one state to jump into more than one state that is not possible forward.

So only one step forward is possible because of the arrival process follows the Poisson process and since we have only one server in the system, the system also decremented by only one level below, therefore this is going to form a birth death process, the reason for this CTMC going to be a birth death process because of the arrival process follows a Poisson process so whatever the assumptions we have it for the Poisson process that is going to be satisfied.

And since we have only one server in the system and he does the service for only one customer at a time after finishing that server - after finishing the customer service then it move into the next service immediately and so on, if the customers are available in the queue, therefore the system goes to the one step one state below by only one move only, it won't move from 2 to 0 or 3 to 1 and so on, therefore this CTMC is a birth death process.

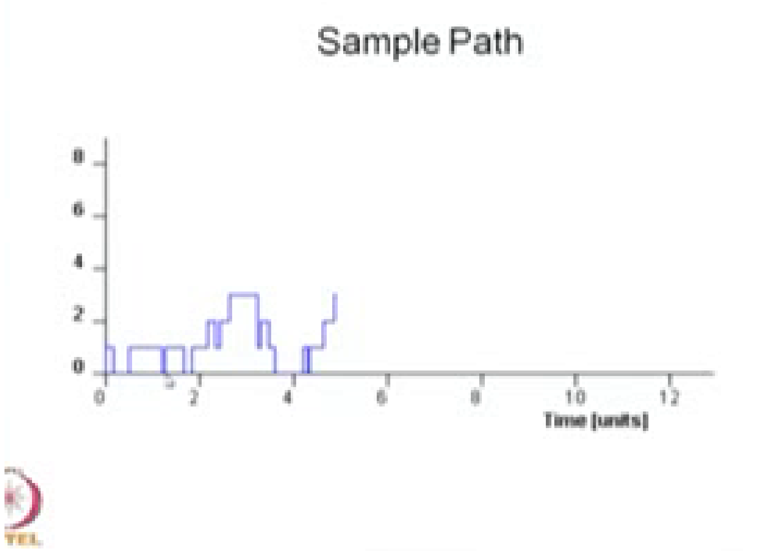
(Refer Slide Time: 07:05)



Therefore I am connecting the CTMC with the MM1 queue in particular the CTMC the birth death process, because of the transitions due to arrival or departure of a customer and only nearest neighbors transitions are allowed, because of the assumptions which we have made, therefore this is going to a continuous time Markov chain with the rate in which the system moves from the state i to i + 1, that rate is lambda.

And the system moves from the state i to i minus 1 that rate is mu, and all other rates are going to be 0 other than the diagonal element and this rates also constant not state dependents rates, therefore this is a birth death process with the birth rates lambda and the death rates mu.

(Refer Slide Time: 08:09)



So this is a sample path suppose a time 0, one customer in the system then it services over then the second customer enter into the system, now the number of customers in the system is 1 and so on, so that means this duration is the service time for the first customer and from this point to this point that is a inter arrival time of the second customer enter into the system.

And from this time point to this time point that is the service time for the second customer which is independent of the service time for the first customer and this is the time point the second customer enter and this is the time point in which the third customer enters therefore the inter arrival time is from this point to this point and so on, so this is a dynamics of a number of customers in the system over the time.

Therefore this stochastic process is a discrete state continuous time stochastic process satisfying the Markov property, therefore this is a continuous time Markov chain. So later I am going to simulate the MM1 queuing model using some simulation technique. So the conclusion is the under the underlying stochastic process for the MM1 queuing model is a birth death process N of t is a stochastic process, so this stochastic process is a birth death process.

Therefore now we are going to discuss the stay - stationary distribution time dependent probabilities and so on.

(Refer Slide Time: 10:01)

Stationary Distribution

$$\pi = (\pi_0, \pi_1, \cdots) \; ; \; \pi_i \geq 0 \; ; \; \sum_i \pi_i = 1$$

$$\pi Q = 0 \qquad \pi_i = P[N = i]$$

$$N = \lim_{t \to \infty} N(t)$$

$$0 = -\lambda \pi_0 + \mu \pi_1$$

$$0 = \lambda \pi_{i-1} - (\lambda + \mu)\pi_i + \mu \pi_{i+1}, \; i \geq 1$$

$$\pi_1 = \frac{\lambda}{\mu} \pi_0$$

$$\pi_{i-1} = \frac{\lambda}{\mu} \pi_i = \frac{\lambda^{i-1}}{\mu^{i-1}} \pi_0 \; ; \; i = 1, 2, \cdots$$

So how to find the stationary distribution, solve pi Q is equal to 0, pi is the vector consists of a pi i's, where pi i's are nothing but what is the probability that N customers in the system what is the probability that i customers in the system in a long run, so that long run is defined in this way the N of t is a stochastic process as t tends to infinity, the number of customers in the system in a long run that is going to be the N.

And pi is nothing but a probability that N i customers in the system in a longer run, so now we are going to solve pi Q is equal to 0 with the normalized equation summation of pi i is equal to 1, so once you frame the equation you will get a pi 1, in terms of pi naught and pi i minus 1 in terms of a first pi i, then substitute recursively you will get in terms of pi naught.

So since it is a homogeneous equation you will get all pi i's in terms of pi naught, so use the normalizing equation summation of pi i is equal to 1, you will get pi naught.

(Refer Slide Time: 11:10)

Take $\rho = \frac{\lambda}{\mu}$

Then,

$$\pi_0 = 1-\rho$$

$$\pi_n = (1-\rho)\rho^n \quad ; \quad \rho < 1 \text{ (stable system)}, \quad n = 1, 2, \ldots$$

ρ : offered load (traffic intensity)

ρ : server utilization

So the pi naught is equal to 1 minus rho, where rho is lambda by mu and since I am relating this stochastic process with a birth death process with a infinite capacity, if you recall the stationary distribution exist as long as the denominator of pi naught that series converges, so that will converge only if lambda by mu is less than 1, if lambda by mu is greater than or equal to 1 then that denominator diverges accordingly you won't get the stationary distribution.

So to have a stationary distribution you need rho has to be less than 1, that also you can intuitively say whenever system is stable that is corresponding to rho is less than 1 in that you will have a stationary distribution that means in a longer run this is the proportion of the time the system will be empty and the pi n is nothing but the n customers in the system in a longer run, that is a 1 minus rho times rho power n, where rho is less than 1.

This rho can be visualized as the offered load also, because the rho is nothing but the mean arrival rate and the mu is a mean service rate and this ratio will give the offered load and 1 minus pi naught that is the probability that the system is non-empty and that is nothing but the server utilization.

Server utilization is nothing but what is the probability that the server is busy, the server will be busy as long as the system is non-empty, so the rho is the server utilization that can be obtained in the from this formula and in a longer run the server utilization is rho.