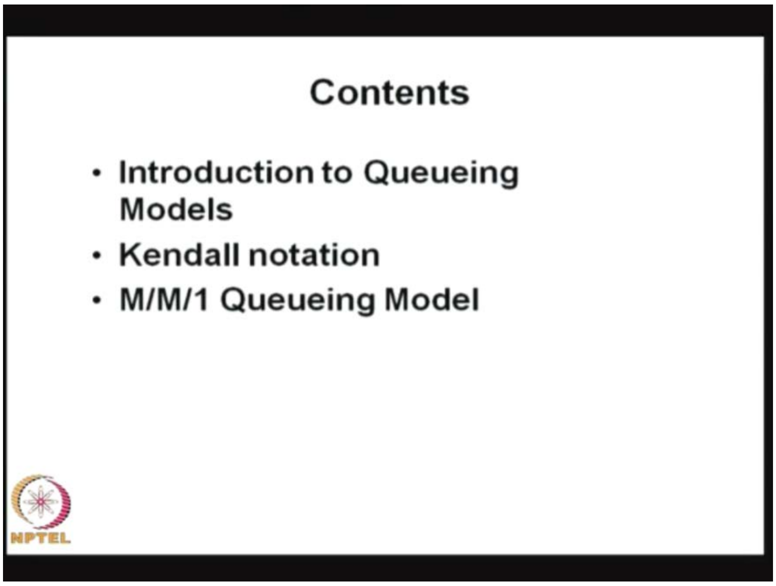
Stochastic Processes
Prof. Dr. S. Dharmaraja
Department of Mathematics
Indian Institute of Technology, Delhi

Module - 5
Continuous-time Markov Chain
Lecture - 4
M/M/1 Queueing Model

(Refer Slide Time: 00:34)

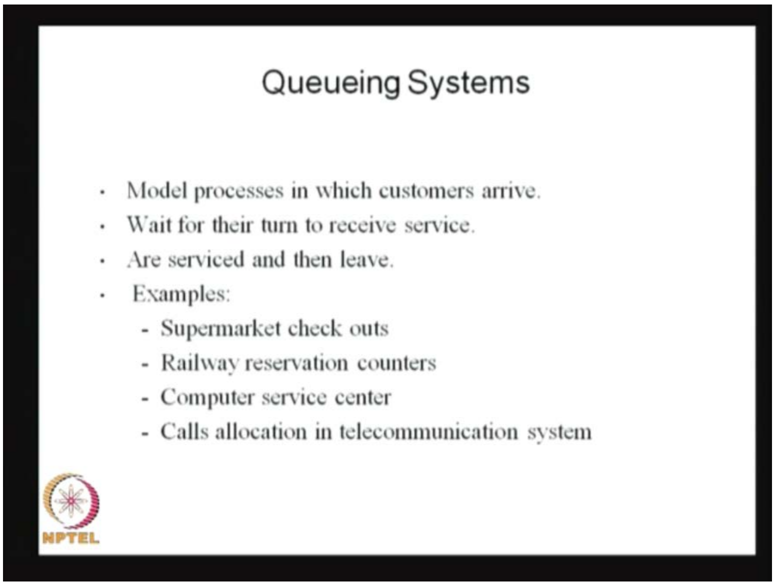


Contents

- Introduction to Queueing Models
- Kendall notation
- M/M/1 Queueing Model

Stochastic process; this is module 5, lecture 4; M M 1 queueing model. In this talk I am going to discuss the queueing models. So, for that I am going to give the introduction to the queueing models. Then I am going to discuss the Kendall notation then followed by that the simplest queueing model M M 1 queue will be discussed. And this is going to be the applications of a continuous time Markov chain in queueing models. So, in this lecture I am going to discuss only the simplest queueing model M M 1 queues.

(Refer Slide Time: 01:10)

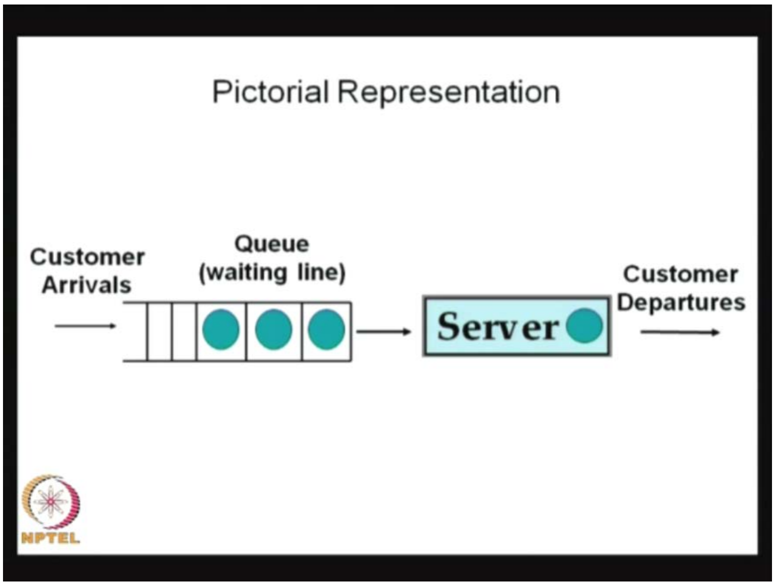


Queueing Systems

- Model processes in which customers arrive.
- Wait for their turn to receive service.
- Are serviced and then leave.
- Examples:
  - Supermarket check outs
  - Railway reservation counters
  - Computer service center
  - Calls allocation in telecommunication system

So, how one can define the queueing system? You can see many examples in which whenever you go to super market to get some items or you see the railway station counters or you can see the computer service center many pc's are there, and printers and so on. So, how the queueing system is created? And also you can see the examples in the calls allocation in telecommunication systems. In all those examples, you can see something is getting served, and leave the system.

(Refer Slide Time: 01:55)



Pictorial Representation

Customer Arrivals

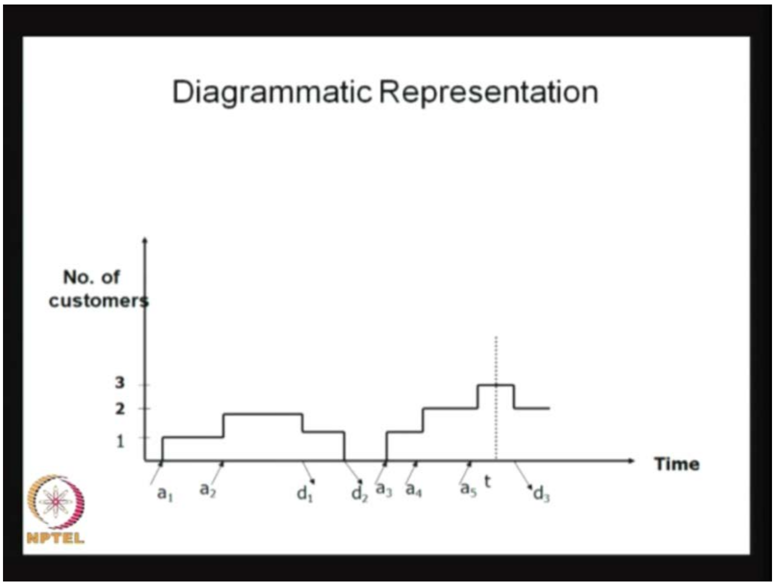
Queue (waiting line)

Server

Customer Departures

We can give the queueing system. We can represent the queueing system in the pictorial form some customers are coming into the system and waiting for their service. Once the service is over then they departure from the system. So, this is the way one can visualizes the queueing system in a pictorial form.

(Refer Slide Time: 02:26)



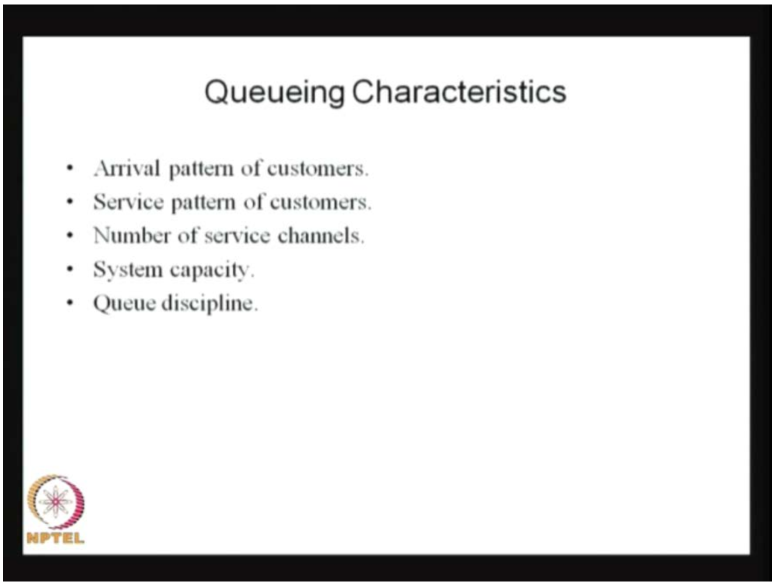
Diagrammatic Representation

This is a diagrammatic representation, and x axis is the time, and y axis is the number of customers in the system. Suppose, at time a 1, the first customer enter into the system then the number of customers in the system is incremented by 1. The customer who enters the system is getting the service during his service time. The next customer enter the system that with the time point a 2. Therefore, now the number of customers in the system is 2, going on, at this time point the first customer service is over.

So, he departure from the system that is d 1, the time point in which the first customer service is over. Now, the number of customers in the system is 1. The time pointer t suffix 2, the second customer service also gets over. Now, the number of customers in the system is 0. The third customer enters at the time point a 3. So, during this interval the system was empty.

So, like that the system is keep increasing whenever one customer enter into the system and decreasing by 1 whenever the service is completed. So, this is the diagrammatic representation of a any queueing system here I made the assumption which very simplest one, only one customer entering to the system. And only one customer is getting served and leave the system and so on. So, this is the simple way of simple diagrammatic representation of the queuing system.

(Refer Slide Time: 04:27)



Queueing Characteristics

- Arrival pattern of customers.
- Service pattern of customers.
- Number of service channels.
- System capacity.
- Queue discipline.

So, to define the queueing system you need a few important characteristics using that one can easily frame the queueing system. So, for that you need the first information that is arrival pattern of customers how the customers are entering into the system? How frequently whether the customers are coming in a very constant interval of time? Or in a random fashion if it is constant then we say the inter arrival time is deterministic. If the customers are entering into the system with the inter arrival time that is some random variable. Then you should know what is the distribution of inter arrival time?

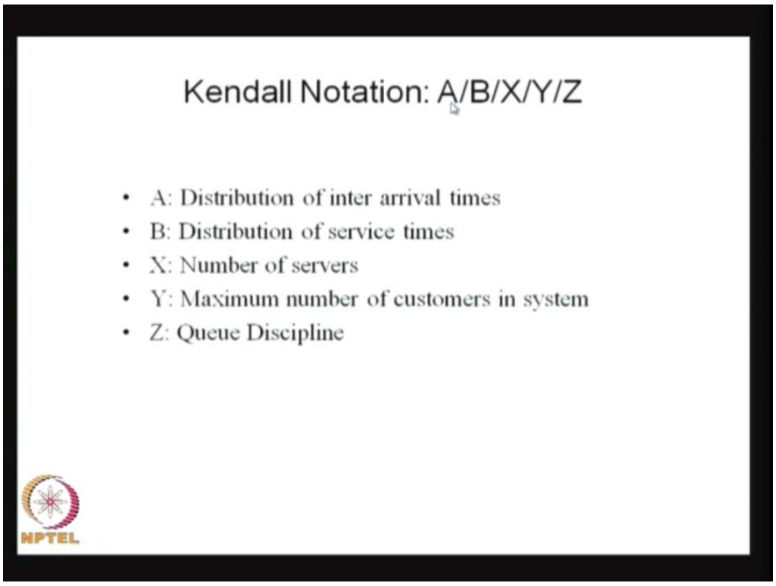
So, this information is needed to define the queueing system, the arrival pattern that includes whether it is deterministic or probabilistic. If it is a deterministic then what is the inter arrival time, the constant time? If it is a probabilistic then what is the distribution and so on. Similarly, after the customers entering into the system, you should know how the service takes place, whether the service time for each customer who enter into the system is it a constant or random. If it is a constant amount of service for each customer then what is a time? How much time it takes for each service?

If it is a probabilistic then what is the distribution of service time. Then the third important information or the characteristics is number of servers in the system. How many service channels are available to do the service whether you have only one server in the system or more than one or countably infinite numbers. So, according to that the queueing system may vary. So, the third information is number of servers in the system.

The forth information that is system capacity, whether the capacity is a finite one or infinite capacity. Accordingly the number of customers in the system may go maximum the finite capacity or it may infinite number of customers can be it in the system to get the service. Therefore, the system capacity is also important characteristic. The forth one queueing discipline, when the customers entering into the system whether they are getting severed or whether they are placed in a first come, first order or first come last service or random fashion or priority based and so on. So, the queueing discipline also important to know the, how the queuing system is at any time, to know the dynamics number of customers in the system. And you should know how the queueing discipline is taken care.

Similarly, the service discipline also how the service is also takes place during the picking the customers for the service. So, these are all the minimum important information to characterize the queueing system; one is arrival pattern, second is a service pattern. And the third is a number of servers, the forth is the capacity of the system, and the service discipline or queueing discipline.

(Refer Slide Time: 08:10)



Kendall Notation: A/B/X/Y/Z

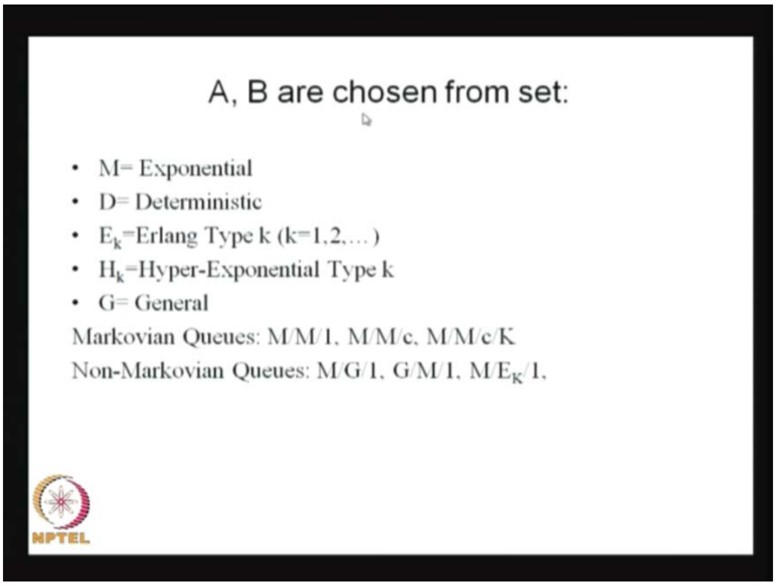
- A: Distribution of inter arrival times
- B: Distribution of service times
- X: Number of servers
- Y: Maximum number of customers in system
- Z: Queue Discipline

So, based on that the Kendall made a notation, and that notation is called a Kendall notation .The Kendall notation consist of capital A that letter A slash capital B slash capital X slash capital Y, and capital Z. So, the possible values we are going to assign for A B X Y Z accordingly one can define the queuing system. And each letter is corresponding to some important characteristic of the queueing system.

A denotes the arrival pattern information, here the A denotes the distribution of inter arrival time, the letter corresponding to the capital A. The second one capital B whatever the letters you are going to assign for the second one that denotes the distribution of the service time. The way I have said the characteristic the first one is arrival pattern, second one is service pattern, and so on the same way we have given the Kendall notation. So, the capital A is for the letter whatever the letter you are going to assign for capital A that is for the distribution of inter arrival time.

And B is for the service time distribution. The third one X whatever the number you are going to write that is the number of servers in the system. The forth one what is the capacity of the system? The fifth one what is the queueing discipline? Whether it is a first come first served, last come first served priority random and so on. Now, I am going to give what are all the different possible values for the, this letters.

(Refer Slide Time: 10:14)



A, B are chosen from set:

- M= Exponential
- D= Deterministic
- E_k=Erlang Type k (k=1,2,…)
- H_k=Hyper-Exponential Type k
- G= General

Markovian Queues: M/M/1, M/M/c, M/M/c/K

Non-Markovian Queues: M/G/1, G/M/1, M/E_K/1,

The first two, A is for the distribution of inter arrival time; B is for the distribution of service time. The both can be chosen from this letters. If you write M in the first place that means the inter arrival time is exponentially distributed. Even though it is exponentially distributed, we use the letter M because of exponential distribution satisfies the memory less property or Markovian property so, to denote that we use the letter M. So, whenever you write M in the place of A or the second place B, then that means the inter arrival time is exponentially distributed or service time is exponentially distributed respectively.

Suppose, you write the letter D in the place of A or B that means that distribution is a deterministic. That means it is going to take it is not a probabilistic, it is takes a constant amount of time whether you placed it in the first or second accordingly. So, it is going to be a constant amount of time going to take for the inter arrival time or service time whenever you place it in A or B respectively. Similarly, if you use the letter E suffix k that means it is erlang distribution of type k or we can say erlang distribution of stage k that can be 1 2 and so on. That means the inter arrival time is a erlang distributed in the stage k if you place it in the first letter. Similarly, H suffix k means hyper exponential distribution of a type k, whenever you have a inter arrival time is other than exponential deterministic, and so on.

So, usually other than exponential you can use the letter G. G means general distribution; general distribution is also it is a known distribution. The only thing is it is other than exponential distribution. So, either you can use the letter M D E k H k or G. So, G can be other than M itself in the usual or in general form. So, known distribution that other than exponential we use the letter called G for general distribution. So, these are all the possible values for the A and B, whereas the third one is the number of servers in the system. And the fourth one is the capacity of the system. And the fifth one is the queueing discipline.

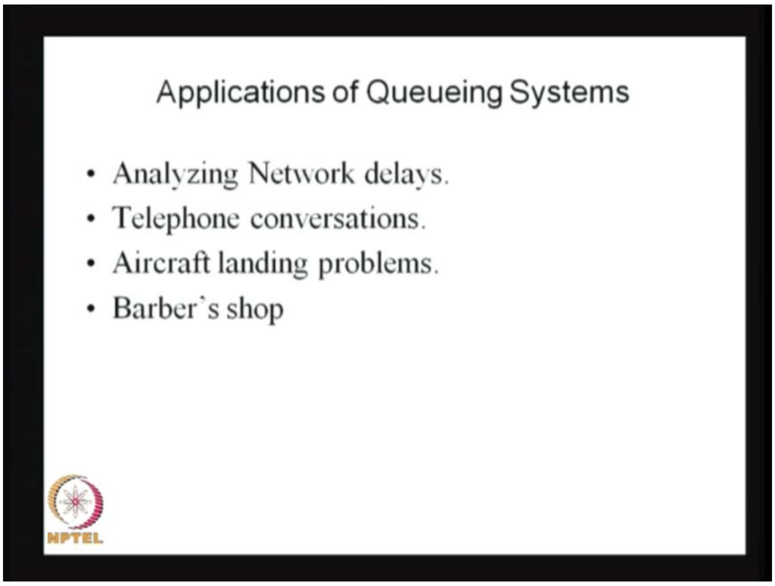
The default discipline is first come first out. Therefore, no need to write the fifth information and the sixth information is also there. What is the population of the customers who are entering into the system? The default population is infinite. That means from infinite source the customers are entering into the system that is the sixth information. As long as we would not write, as long as the system in which the population is infinite, as well as the queueing discipline is first come first served, then we would not write. So, we write only the first four information, that is inter arrival time, distribution; second one is a service time distribution. The third one is number of servers, and the forth information is capacity of the system.

So in these examples, the inter arrival time and service time both are exponentially distributed, by default they are independent also. And the third letter denotes number of servers in the system. So, here only one server in the system, here c means it can be greater than or equal to 1 that is a multi-server system. And forth letter k means capacity of the system.

Suppose, we did not write fourth information here that means it is a infinite capacity system, and this is also infinite capacity system. And since, the inter arrival time, and the service time are exponentially distributed. This model is called the Markovian queues, because it satisfies the Markov property whereas, non Markovian queues either service time or the inter arrival time can be a non-exponential distribution or non-exponential distribution in default you can use the letter in general distribution.

So, whenever G comes in first place or the second place, then we use non, then we say it is a non non Markovian queues. And if the forth letter is missing that means it is a infinite capacity system.

(Refer Slide Time: 15:24)

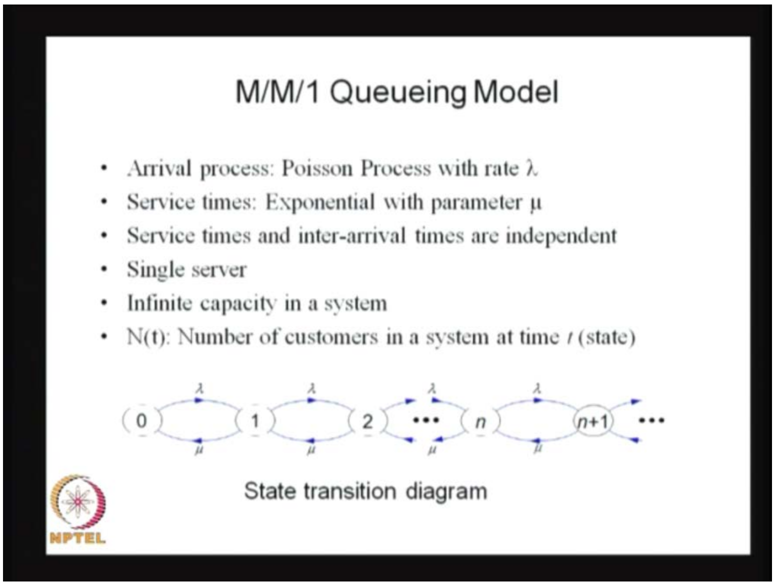


Applications of Queueing Systems

- Analyzing Network delays.
- Telephone conversations.
- Aircraft landing problems.
- Barber's shop

There are many applications of queueing system we are going to discuss the abstract queuing system in the further lecture. The easiest, or the simplest queueing model that is Markovian queuing model that is M M 1 queueing model. Later we are going to relate with the birth death process also. In the M M 1 queueing model, the inter arrival time is exponentially distributed as I discussed the Poisson process in the previous lecture.

(Refer Slide Time: 15:35)



M/M/1 Queueing Model

- Arrival process: Poisson Process with rate λ
- Service times: Exponential with parameter μ
- Service times and inter-arrival times are independent
- Single server
- Infinite capacity in a system
- N(t): Number of customers in a system at time *t* (state)

State transition diagram

Whenever you have arrival follows a Poisson process then the inter arrival time follows exponential distribution, and or independent also. So, here the first information is arrival process follows Poisson process with the intensity or rate lambda that means the inter arrival times are independent. And each one is exponential distributed with the parameter lambda.

The second information that is service time; service times are exponentially distributed with the parameter mu. And the service times are independent for each customer and that also independent with the arrival process that means there is no dependency over the arrival pattern with the service pattern. Service times and the inter arrival times are independent then the third information, only one server in the system that is the queuing system in which only one server.

And the forth information is missing that means it is a default it is a infinite capacity model. Now, our interest is to find out the behavior of a queuing system or the behavior of the number of customer 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 in the inter arrival time is exponentially distributed, and the service times are exponentially distributed. The memory less property 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 times also independent for each customer. Therefore, this stochastic process satisfies the memory less property at all-time points. Therefore, this 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 as 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 a discrete state continuous time stochastic process satisfying the Markov property based on this assumptions. Therefore, N of t is a Markov process, since the state space is discrete therefore, this is a Markov chain. Therefore, this is a continuous time Markov chain. Therefore, N of t is a c t m c. So, one can write the state transition diagram for the for this c t m c 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.

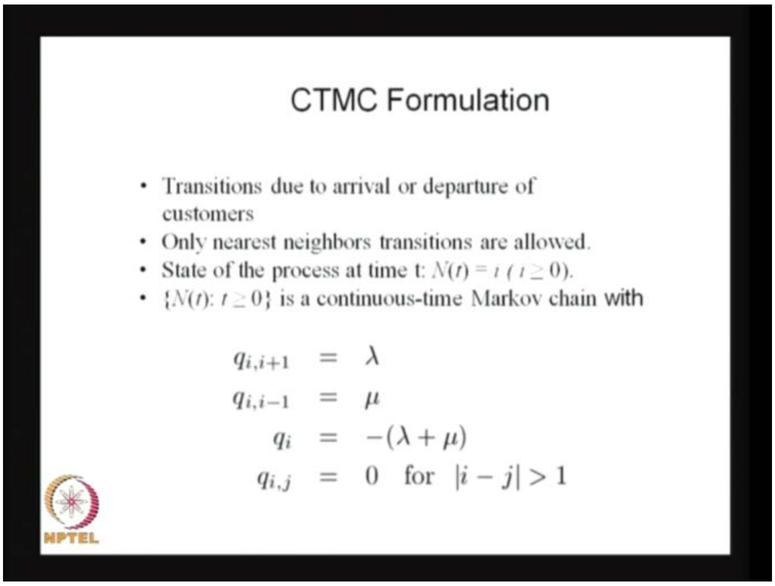
Since it is a M M 1 queue model, queuing model therefore, whenever the system is in the whenever the system is in the state 0 by the inter arrival time which is exponentially distributed. The number of customers in the system will be incremented by 1 therefore, that rate will be lambda or the system moving from the state 0 to 1 it spends exponentially distributed amount of time here before moving into the state 1. Once 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 the parameter mu.

Therefore, the system goes from the state 1 to 0 with the 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 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 c t m c is going to be a birth death process because of the arrival process follows 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 moves into the next service immediately and so on, if the customers are available in the queue. Therefore, the system goes to the one state one step below by only one move only, it would not move from 2 to 0 or 3 to 1 and so on. Therefore, this c t m c is a birth death process.

(Refer Slide Time: 22:29)



CTMC Formulation

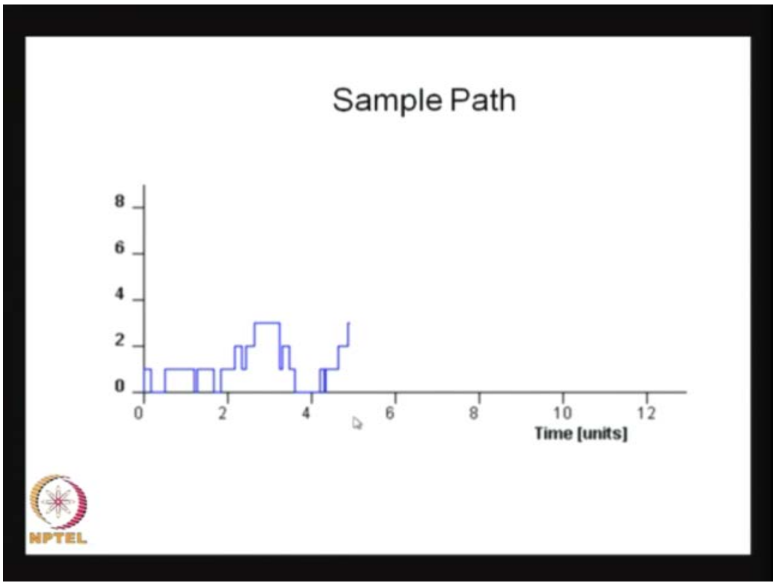
- Transitions due to arrival or departure of customers
- Only nearest neighbors transitions are allowed.
- State of the process at time t: $N(t) = i$ $(i \geq 0)$.
- $\{N(t): t \geq 0\}$ is a continuous-time Markov chain with

$$\begin{aligned} q_{i,i+1} &= \lambda \\ q_{i,i-1} &= \mu \\ q_i &= -(\lambda + \mu) \\ q_{i,j} &= 0 \quad \text{for} \quad |i-j| > 1 \end{aligned}$$

Therefore, I am connecting the c t m c with the M M 1 queue in particular c t m c with the birth death process. Because of the transition due to arrival or departure of a customer, and only nearest neighbor’s transition are allowed because of the assumption which you 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 plus 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 the state dependent rates. Therefore, this is birth death process with the birth rates lambda, and death rates mu.

(Refer Slide Time: 23:32)

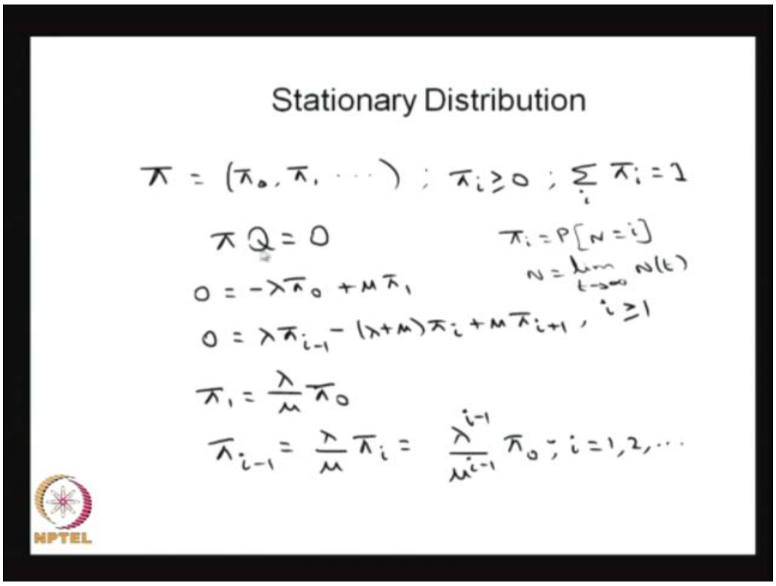


Sample Path

So, this is a sample path suppose at time 0, one customer in the system then service is the over. Then the second customer enters into the system. Now, the number of customer 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 the inter arrival time of the second customer entering 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 points the second customer enter, and this is the time point in which the third customer enter. Therefore, the inter arrival time is from this point to this point and so on.

So, this is the dynamics of 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 the continuous time Markov chain. So, later I am going to stimulate the M M 1 queueing model using some simulation technique. So, the conclusion is the underlined stochastic process for the M M 1 queueing model is a birth death process. The n of t is a stochastic process. So, this stochastic process is a birth death process. Therefore, now we are going to discuss the stationary distribution time dependent probabilities and so on.

(Refer Slide Time: 25:24)



Stationary Distribution

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

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

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

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

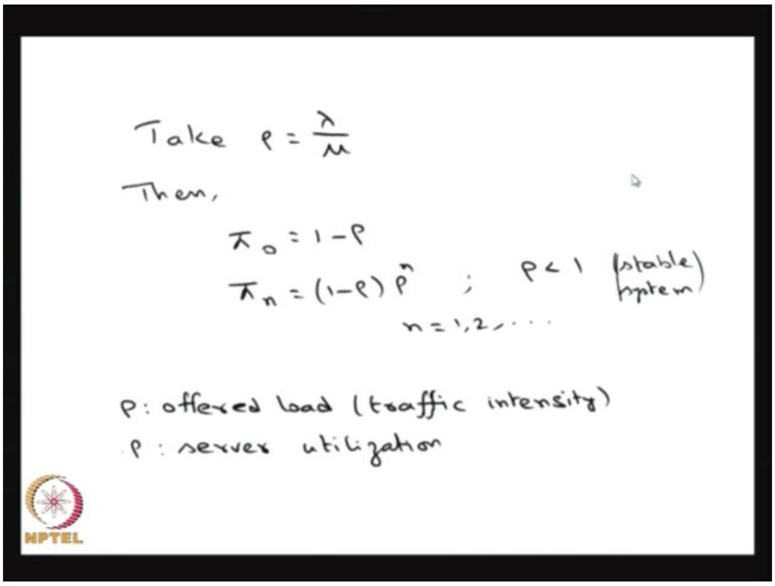
$0 = \lambda \pi_{i-1} - (\lambda + \mu) \pi_i + \mu \pi_{i+1} \, , \; i \geq 1$

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

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

So, how to find the stationary distribution? Solve by q is equal to 0, pi is the vector consist of a pi i’s where pi i’s are nothing but, what is the probability that n customers in the system? What is a 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 probability that N that i customers in the system in a longer run.

(Refer Slide Time: 26:33)



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

Then,

$\pi_0 = 1 - \rho$

$\pi_n = (1 - \rho) \rho^n \; ; \quad \rho < 1$ (stable system)

$n = 1, 2, \cdots$

ρ: offered load (traffic intensity)

ρ : server utilization

So, now, we are going to solve pi q is equal to 0 with the normalized equation summation of pi is equal to 1. So, once you frame the equation you will get pi 1 in terms of pi naught. And pi i minus 1 in terms of 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

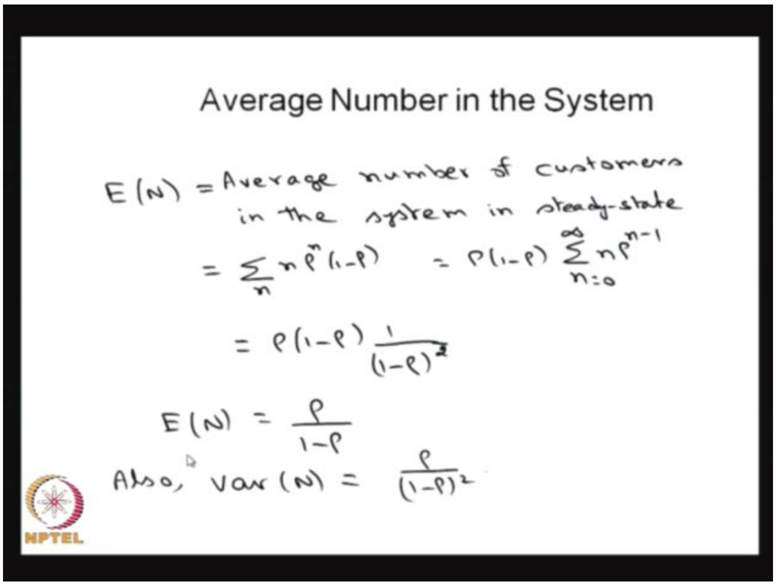
So, use the normalizing equation summation of pi i is equal to 1, you will get pi naught. So, the pi naught is equal to 1 minus rho where rho is lambda by mu, and since I am relating this stochastic process with a birth death process with the infinite capacity. If you recall the stationary distribution 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 would not get the stationary distribution.

So, to have stationary distribution you need rho has to be less than 1, that also you can tentatively say whenever the system is stable that is corresponding to rho is less than 1 in that you will have the stationary distribution. That means in a longer run, this is a 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 1 minus rho times rho power n where rho is less than 1. This rho can be visualized as a 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 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 obtain in the, from this formula, and in a longer run the server utilization is rho.

Other than stationary distribution one can find out the average measures also in the system. So, suppose you make a E of n that is nothing but the average number of customers in the system in steady state. Since, you know the probability distribution, substitute p pi hence here. Therefore, n times pi n summation over n that is going to be the average number of customers in the system.

(Refer Slide Time: 28:29)



Average Number in the System

E(N) = Average number of customers in the system in steady-state

$$= \sum_{n} n\rho^{n}(1-\rho) = \rho(1-\rho)\sum_{n=0}^{\infty} n\rho^{n-1}$$

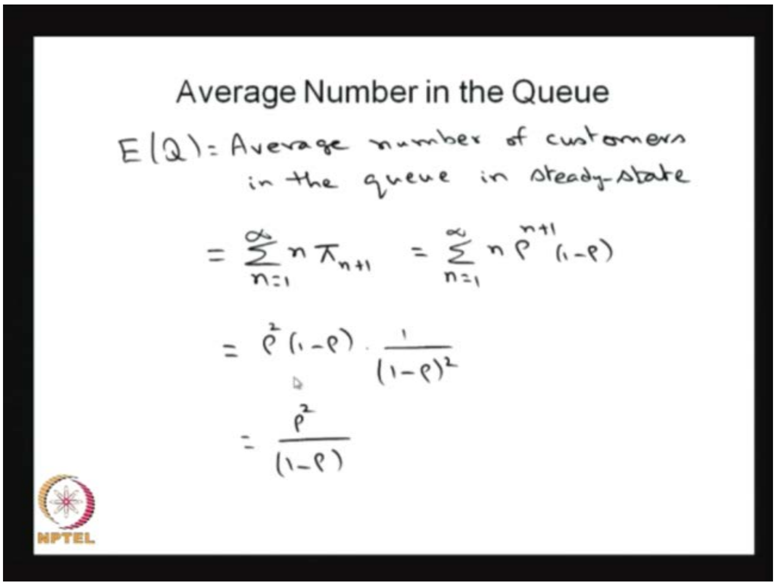
$$= \rho(1-\rho)\frac{1}{(1-\rho)^2}$$

$$E(N) = \frac{\rho}{1-\rho}$$

Also, $\mathrm{Var}(N) = \frac{\rho}{(1-\rho)^2}$

If you do little simplification, you will get rho divided by 1 minus rho where rho is less than 1. So, this is the average number of customers in the system. And also, one can get number variance of the of customers in the system also, for that you have to find out the E of n square then using that formula you can get the variance of n also. So, here we are getting a mean and variance of number of customers in the system in steady state.

(Refer Slide Time: 29:33)



Average Number in the Queue

E(Q): Average number of customers in the queue in steady-state

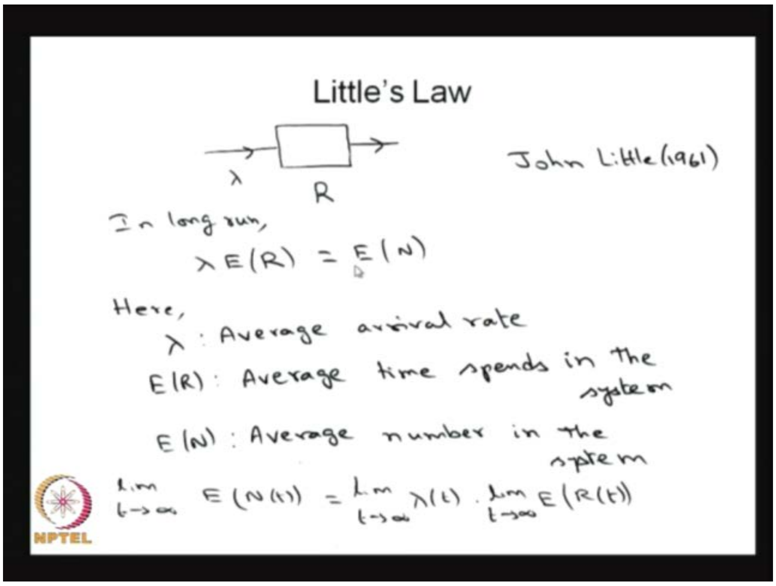
$$= \sum_{n=1}^{\infty} n\pi_{n+1} = \sum_{n=1}^{\infty} n\rho^{n+1}(1-\rho)$$

$$= \rho^2(1-\rho)\cdot\frac{1}{(1-\rho)^2}$$

$$= \frac{\rho^2}{(1-\rho)}$$

Also one can find average number in the queue. So, the letter Q is a random variable. And here, we are finding the expectation of Q that is the average number of customers in the queue. That means before getting the service how many customers in the system? We have only one server in the system. And whenever the service is going on, and all other arriving customers will be queued. That means when n plus 1 customer in the system, n people are in the queue therefore, summation n times pi n plus 1. Do the simplification you will get average number of customers in the queue also. Substitute the pi n plus 1 from the one I have discussed in the stationary distribution.

(Refer Slide Time: 30:27)



Little's Law

John Little (1961)

In long run,

$$\lambda E(R) = E(N)$$

Here,

λ : Average arrival rate

$E(R)$: Average time spends in the system

$E(N)$: Average number in the system

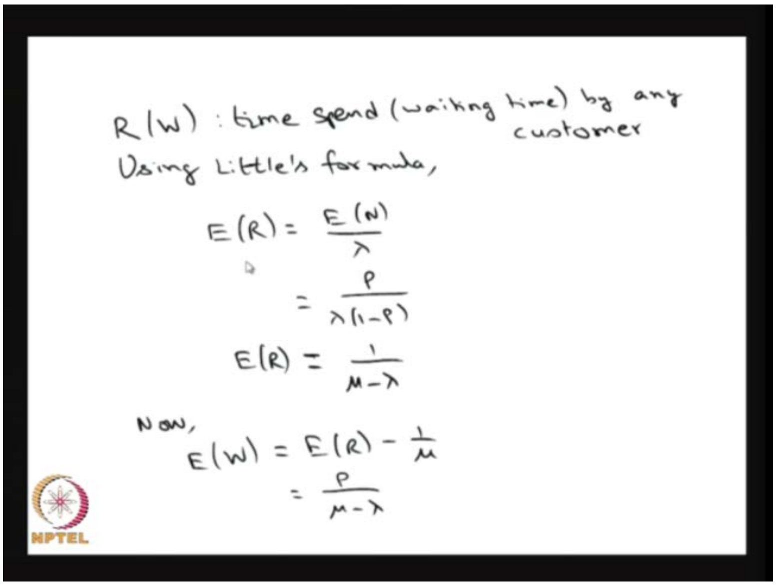
$$\lim_{t \to \infty} E(N(t)) = \lim_{t \to \infty} \lambda(t) \cdot \lim_{t \to \infty} E(R(t))$$

Here I am going to relate the average measures using the little's law. This is proven by John Little 1961. This is valid for any system in which arrival comes into the pattern with the arrival rate lambda, and R is the time spends in the system. And leave the system after the service or whatever the things are over. Then in a longer run, one can say the arrival rate multiplied by the average time spend in the system that is same as average number in the system. So, this relation is valid for whatever be the underlined distribution, whatever be the underlined distribution of the service underlined distribution of the arrival. What it says if you have a system in which the arrival rate is mean arrival rate is lambda.

And the mean time spend in the system is expectation of r. Then that product will give average number in the system. Since, indirectly it says whenever the system has a long run in a stable system, the expectation of average number of customers during the interval 0 to t, as a t tends to infinity that is going to be have a limit expectation of N. And the arrival rate lambda of t that also has the mean arrival rate as a t tends to infinity that is also going to be a sum having a limit constant lambda. And similarly, the average spent by the customers in the system at any time t. And if you make a t tends to infinity that expectation quantity also has limit.

Therefore, we will have lambda times expectation of R is as same as expectation of N. Now, using this little's law I am going to find out the measures for the M M 1 queue model. So, suppose R denotes the time spent in the system by the customer, and W denotes the waiting time by any customer in the system. Then I can use the little's formula the little's law in the previous one. So, if i know the mean arrival rate, and if I know mean number of customers in the system in a longer run, using this I can find out the average time spent in the system. If I know to use the, to use using little's law, if I know the average number of customers in the system longer run. And if I know the arrival rate, then I can find out the average time spent in the system in a longer run.

(Refer Slide Time: 33:32)



R(W) : time spend (waiting time) by any customer

Using Little's formula,

$$E(R) = \frac{E(N)}{\lambda}$$

$$= \frac{\rho}{\lambda(1-\rho)}$$

$$E(R) = \frac{1}{\mu-\lambda}$$

Now,

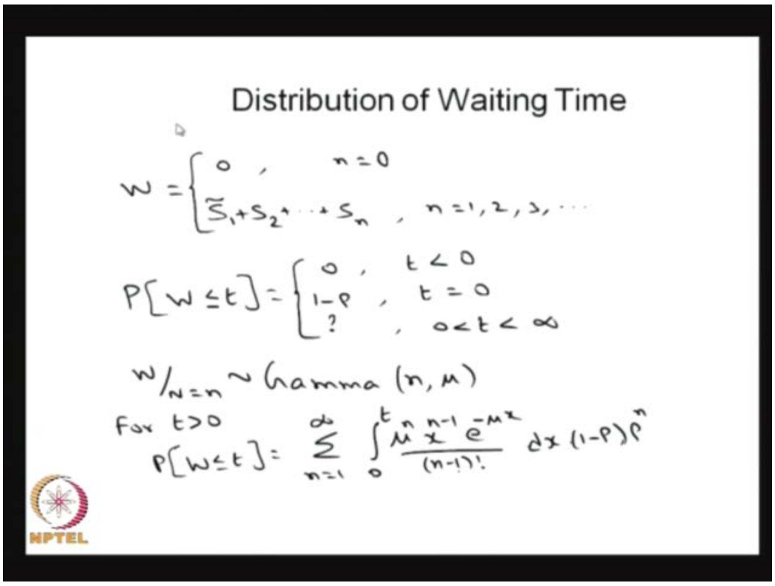
$$E(W) = E(R) - \frac{1}{\mu}$$

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

Similarly, I can once I know the average time spend in the system, if I subtract the average time of my own service, then that is going to be the average time waiting in the queue. So, this is the average time waiting in the queue that is same as average time spend in the system minus my own average service time. The M M 1 queue model, the service time is exponentially distributed with the parameter mu. And therefore, the average is 1 by mu.

So, the difference will give the average time waiting in the queue by any customer. Not only the average measure for the M M 1 queue, one can find out the actual distribution for R as well as W also. Because this is a very simplest Markovian queuing model whereas, for all other models it is little complicated but still one can get it. So, this is the easy model in which one can find out the distribution of the time spend in the system as well as the time or as well as the waiting time by a customer in the queue.

(Refer Slide Time: 34:48)



Distribution of Waiting Time

$$W = \begin{cases} 0, & n = 0 \\ \tilde{S}_1 + S_2 + \cdots + S_n, & n = 1, 2, 3, \cdots \end{cases}$$

$$P[W \leq t] = \begin{cases} 0, & t < 0 \\ 1 - \rho, & t = 0 \\ ?, & 0 < t < \infty \end{cases}$$

$$W/_{N=n} \sim \text{Gamma}(n, \mu)$$

For $t > 0$

$$P[W \leq t] = \sum_{n=1}^{\infty} \int_0^t \frac{\mu^n x^{n-1} e^{-\mu x}}{(n-1)!} dx \, (1-\rho)\rho^n$$

First, let us go for finding out the distribution of waiting time. Waiting time means if no one in the system when you arrive then your waiting time is 0; you are immediately going to cut the service. So, the service time is your time spent in the system. So, usually the time spent in the system is the time of your service plus the time of the waiting time. So, here I am finding the only the distribution of waiting time first. So, whenever the system is 0 your waiting time is 0, whenever no customer in the system the waiting time is 0. Whenever more than or equal to 1 customer in the system then the waiting time is same as the remaining service time for the customer who is under service plus the customers in the queue before you join in the queue.

So, those people service time addition plus the residual or the remaining service time of the customer who the first customer who is under service. So, in this total time is the waiting time whenever the system is non-empty. Whenever the system is empty then the waiting time is 0. Therefore, the W is a random variable either it takes a value 0 or it takes a value greater than 0 based on the time of service of previous n people in the ahead of u. Therefore, W is a mixed random variable which has the probability mass function at 0 as well as density function between the interval 0 to infinity. So, let me try for finding out the c d f of this random variable. So, this c d f is going to be 0 till 0 at 0 it has the c d f 1 minus rho because when the waiting time is 0 that is equivalent of no one in the system.

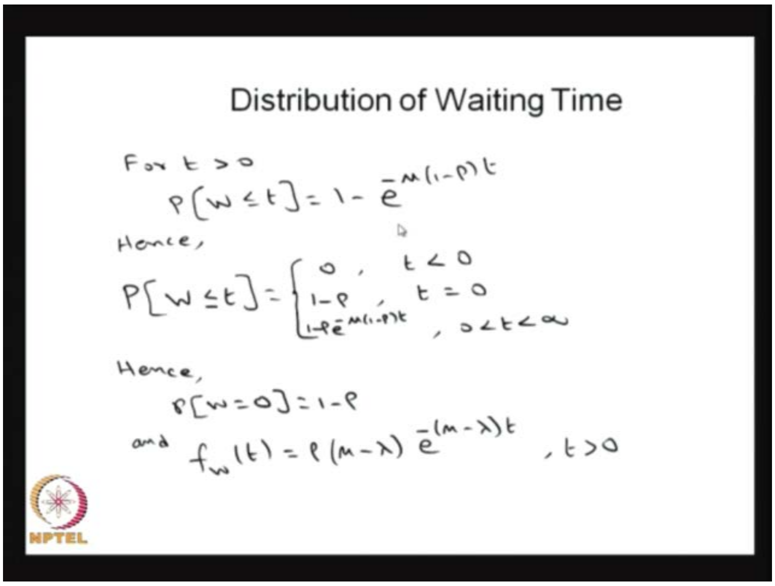
So, in the long run no one in the system that is pi naught, and the pi naught probability is a pi naught probability is 1 minus rho. The system is empty in a longer run that is 1 minus rho. Therefore, the c d f at 0 that is same as 1 minus rho of that is pi naught between the interval 0 to infinity 1 we have to find out the distribution of W. Whenever n customers before you, before you join in the system that conditional distribution the distribution of W given the number of customer in the system is n.

That distribution is nothing but the service time of the n customers. The first customer is a remaining service time. The service time of the first customer is exponential distribution. The residual or remaining service time of the first customer that is also exponential distribution because of memory less property. So, this is exponential distribution. This is the second customer service time that is exponential distribution. And similarly, for the n th customer also the service time is exponentially distributed.

And the way we made assumption all the service times are independent, and each one is exponential distributed with the parameter mu. Therefore, this is a sum of n independent exponential distributed random variables. Therefore, the sum of n exponentials that is going to be a gamma distribution with the parameters n and mu, there are many ways of finding out the distribution. But here I am just explaining through the distribution concept. This is some of n independent exponential distribution therefore, you can conclude it is gamma distribution with the parameter, parameters n and mu. Once you know the conditional distribution our interest is to find out the unconditional one. That means for t is greater than 0 and c d f at the point t that is nothing but what is the conditional density, probability density?

And what is the probability of n customers in the system? That multiplication with the possible n will give the c d f between the intervals 0 to t. So I have density function of a gamma distribution probability density function with the parameters n and mu. And this is the probability density function multiplied, and integration between 0 to t that will give the c d f. And unconditional multiplied by probability of n customers in the system that with the summation that will give the unconditional. Therefore, the c d f is going to be summation n is equal to 1 to infinity integration 0 to t of the probability density function of gamma distribution multiplied by n customers in the system.

(Refer Slide Time: 40:34)



Distribution of Waiting Time

For $t > 0$

$$P[W \leq t] = 1 - e^{-\mu(1-\rho)t}$$

Hence,

$$P[W \leq t] = \begin{cases} 0, & t < 0 \\ 1-\rho, & t = 0 \\ 1-\rho e^{-\mu(1-\rho)t}, & 0 < t < \infty \end{cases}$$

Hence,

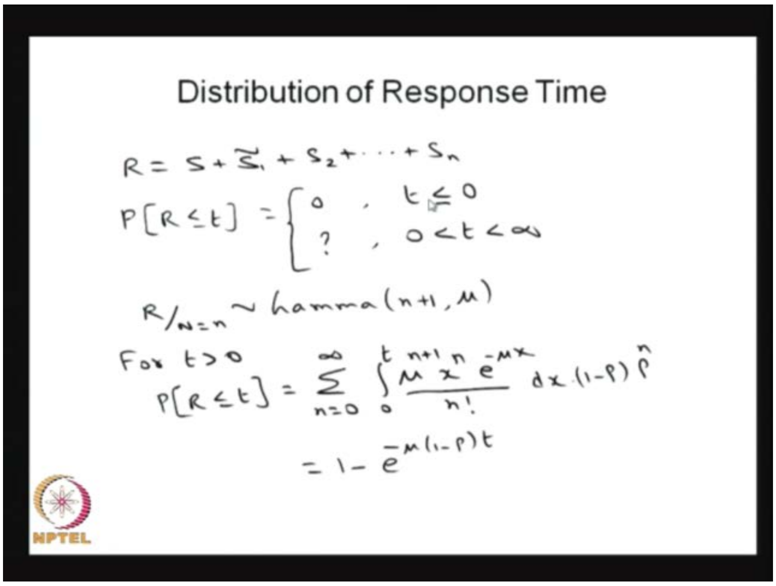
$$P[W = 0] = 1 - \rho$$

and $f_W(t) = \rho(\mu - \lambda) e^{-(\mu-\lambda)t}, \quad t > 0$

If you do simplification, you will get 1 minus e power minus mu times 1 minus rho times t. Therefore, you can substitute here, here I made a mistake. So, here it is multiplied by rho. So, 1 minus rho times e power minus rho, e power minus mu times 1 minus rho that is going to be the. So, once you are getting the c d f we can conclude this is a mixed random variable with the probability mass at 0 is 1 minus rho. And the density function between the interval 0 to infinity that is rho times mu minus lambda times e power minus mu minus lambda times t that is the probability density function for a distribution of waiting time.

Similarly, one can get the distribution of response time also or the total time spend in the system. The total time spend in the system that is nothing but that is a random variable. And the residual service time of the first customer who is in the system plus all the remaining n costumers in the system in the queue plus your own service time. Therefore, here it is not a mixed random variable. This is a continuous random variable, because your service time is a continuous random variable which is exponentially distributed. Therefore, the R is going to be sum of your own service plus the remaining service of the first person in the system if, and so on till the nth customer who is in the queue.

(Refer Slide Time: 41:57)



Distribution of Response Time

$R = S + \tilde{S}_1 + S_2 + \cdots + S_n$

$$P[R \leq t] = \begin{cases} 0 & , \ t \leq 0 \\ ? & , \ 0 < t < \infty \end{cases}$$

$R/_{N=n} \sim Gamma(n+1, \mu)$

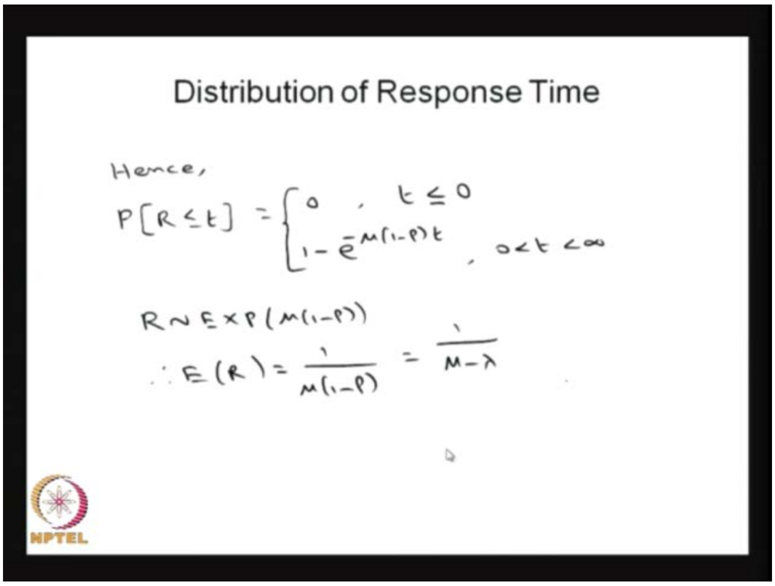
For $t > 0$

$$P[R \leq t] = \sum_{n=0}^{\infty} \int_0^t \frac{\mu^{n+1} x^n e^{-\mu x}}{n!} dx \, (1-\rho) \rho^n$$

$$= 1 - e^{-\mu(1-\rho)t}$$

Therefore, this is a c d f of the random variable or here also one can argue when n customer in the system before him who enters into the system that is sum of a exponential independent random variable and so on. Therefore, this is going to be a gamma distribution with the parameters n plus 1 mu. And for t greater than 0 find out the c d f using the first conditional then unconditional multiplied by 1 minus rho times rho power n summation over n is equal to 0 to infinity because there is a possibility no one in the system or one customer, two customers and so on. Therefore, the running index is 0 to infinity. Do the simplification, you will get 1 minus e power minus rho times this.

(Refer Slide Time: 43:38)



Distribution of Response Time

Hence,

$$P[R \leq t] = \begin{cases} 0 & , \ t \leq 0 \\ 1 - e^{-\mu(1-\rho)t} & , \ 0 < t < \infty \end{cases}$$

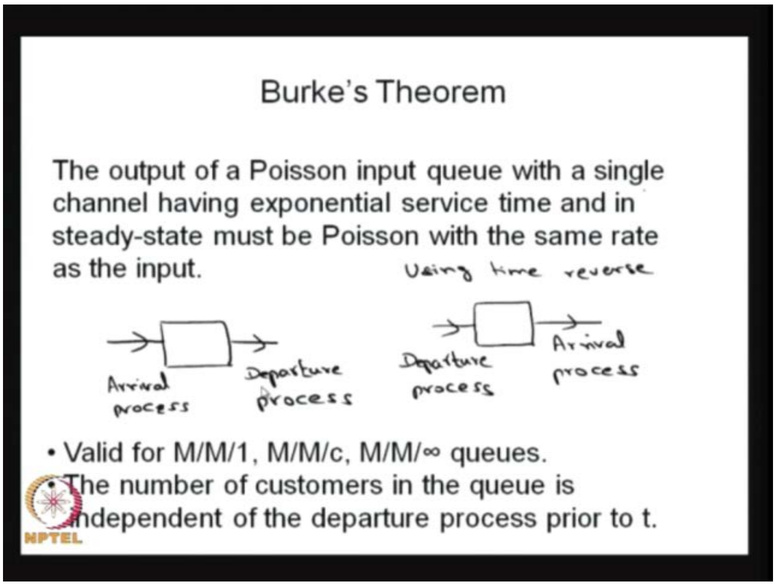
$R \sim EXP(\mu(1-\rho))$

$\therefore E(R) = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu - \lambda}$

Therefore, you can substitute here. And if you see the c d f is same as the c d f of exponential distribution with the parameter that is mu times 1 minus rho. Therefore, you can conclude the total time spend in the system is exponentially distributed with the parameters mu times 1 minus rho.

If you find out the average time that is going to be 1 divided by the parameter that is this the same thing you got it in the average response time from the little's formula using once you know the value of lambda and expected number in the system using little's law you got expectation of time spend in the system that is same result. So, here we are getting first finding the distribution of time spent in the system or response time, then we are finding the average time.

(Refer Slide Time: 44:37)



Burke's Theorem

The output of a Poisson input queue with a single channel having exponential service time and in steady-state must be Poisson with the same rate as the input.

- Valid for M/M/1, M/M/c, M/M/∞ queues.
- The number of customers in the queue is independent of the departure process prior to t.

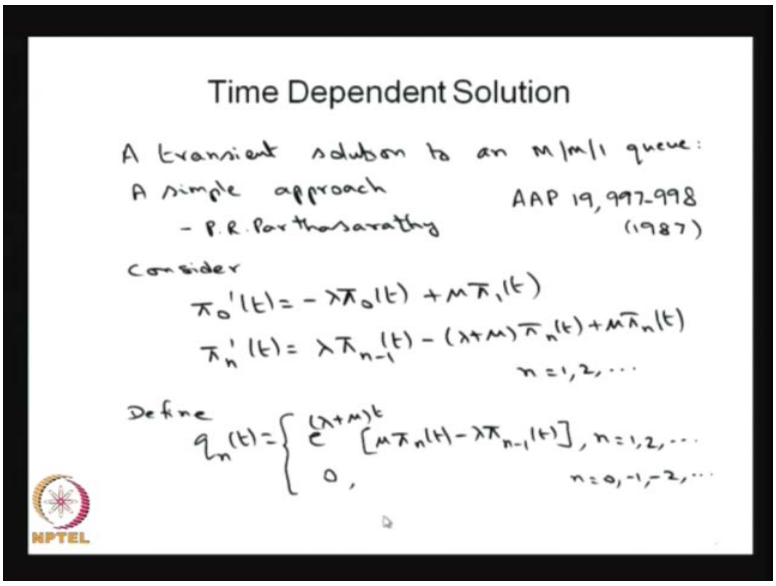
Here I am giving the concept of output process. The arrival follows the Poisson process for the M M 1 queue and the service is exponentially distributed which is independent of the arrival process. And the customers leave the system. Now, the question is what is a distribution of the departure process? That means, what is the inter departure time? After first customer leaves how much time it takes for the second customer leaves the system.

Then the third customer how much time it takes for the inter departure time? And therefore, what is the distribution of the departure process? That is given by the Burke's theorem. The output of the Poisson input queue with the single channel having exponential service time, and in steady state must be a Poisson with the same rate as the input. So, whenever you have a system in which the arrival follows a Poisson. And whenever system has single channel, and the service time is exponential distributed in a longer run the departure process is also going to be a Poisson process.

And the rate will be the same rate as the arrival process. So, this can be proved, but here I am giving the interpretation using the time reverse process. Because in a steady state this module is going to satisfies the time reverse module therefore, the stationary distribution exist. And if you make a, this M M 1 queueing model, the underlying birth death process satisfies the time reversibility equation. Therefore, using the time reverse you can conclude the departure process you can reverse it, and that is going to be independent of the arrival process.

And this is also going to be again Poisson process. So, using the time reverse concept one can prove the departure process is independent of the arrival process. And departure process is also Poisson process with the same rate as the arrival rate. And even though I said it is a single channel having exponential service time. And this is valid for M M 1 queue in the multi-server Markovian queue as well as infinite server Markovian queue also.

(Refer Slide Time: 47:47)



Time Dependent Solution

A transient solution to an M/M/1 queue:
A simple approach
- P.R. Parthasarathy

AAP 19, 997-998
(1987)

Consider

$$\pi_0'(t) = -\lambda\pi_0(t) + \mu\pi_1(t)$$

$$\pi_n'(t) = \lambda\pi_{n-1}(t) - (\lambda+\mu)\pi_n(t) + \mu\pi_n(t)$$

$n = 1, 2, \cdots$

Define

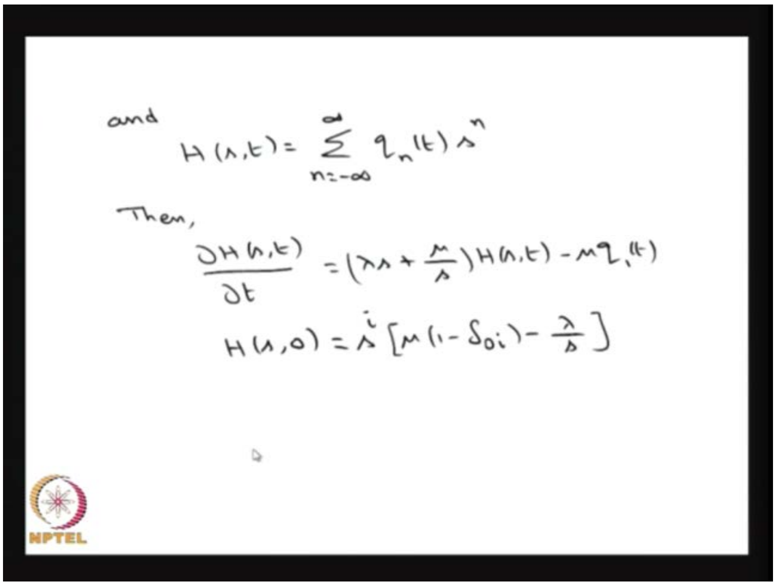
$$q_n(t) = \begin{cases} e^{(\lambda+\mu)t}\left[\mu\pi_n(t) - \lambda\pi_{n-1}(t)\right], & n = 1, 2, \cdots \\ 0, & n = 0, -1, -2, \cdots \end{cases}$$

So, all those thing all those models can be combined with the single channel having exponential service time, whether it is single server or multi server or infinite server this result holds good. And the next result is the number of customers in the queue it is independent of the departure process prior to it that is also satisfies.

Now, we are giving the time dependent solution of a M M 1 queue. There are many more methods to find out the time dependent solution for a M M 1queue. It started with the spectral method and (( )) method, and also with the difference sequence method. Like that there are many more methods in the literature to find out the time dependent solution. And here I am presenting the time dependent solution by P. R Parthasarathy. And this work is appeared in the advance applied probability volume number 19, 1987. So, in this paper they have, he has consider the system of difference differential equation that it nothing but the forward kolmogorov equation.

And making a simple function q n of t that is a difference of pi n's with the multiplication e power lambda plus mu of t. So, once you use this definition once you convert the this system of difference equation with q n of t.

(Refer Slide Time: 48:57)



and

$$H(s,t) = \sum_{n=-\infty}^{\infty} q_n(t)\, s^n$$

Then,

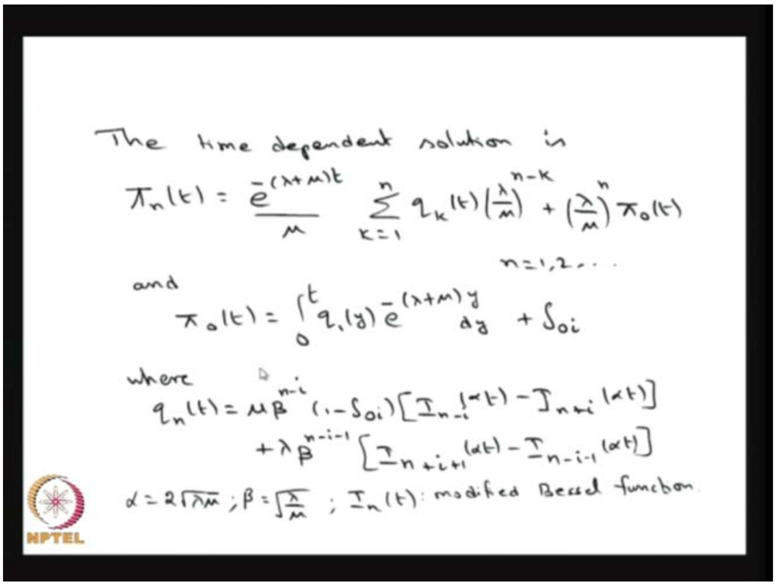
$$\frac{\partial H(s,t)}{\partial t} = \left(\lambda s + \frac{\mu}{s}\right) H(s,t) - \mu q_1(t)$$

$$H(s,0) = s^i \left[\mu(1-\delta_{0i}) - \frac{\lambda}{s}\right]$$

By making a proper generating function that is of the form n, n is equal to minus infinity to infinity, q n of t times s power n. Therefore, this is sort of generating function in terms of q n of t where q n of t is for n is equal to 1 to infinity. This is of difference of mu times pi n minus lambda times pi n minus 1, multiplication e power this function.

And for n is equal to 0 minus 1 2 and so on 0. Therefore, you have a generating function so you can convert the whole difference differential equation in terms of pi n into the one partial differential equation with the initial condition also changes. Because if you assume that i customers in the system at time 0. And this is going to be initial condition for a function H of s comma t at t equal to 0. Now, the question is you have to solve this equation with this initial condition; this p t e using this initial condition.

(Refer Slide Time: 50:13)



The time dependent solution is

$$\pi_n(t) = \frac{e^{-(\lambda+\mu)t}}{\mu} \sum_{k=1}^{n} q_k(t)\left(\frac{\lambda}{\mu}\right)^{n-k} + \left(\frac{\lambda}{\mu}\right)^{n} \pi_0(t)$$

$n = 1, 2, \ldots$

and

$$\pi_0(t) = \int_0^t q_1(y)\, e^{-(\lambda+\mu)y}\, dy + \delta_{0i}$$

where

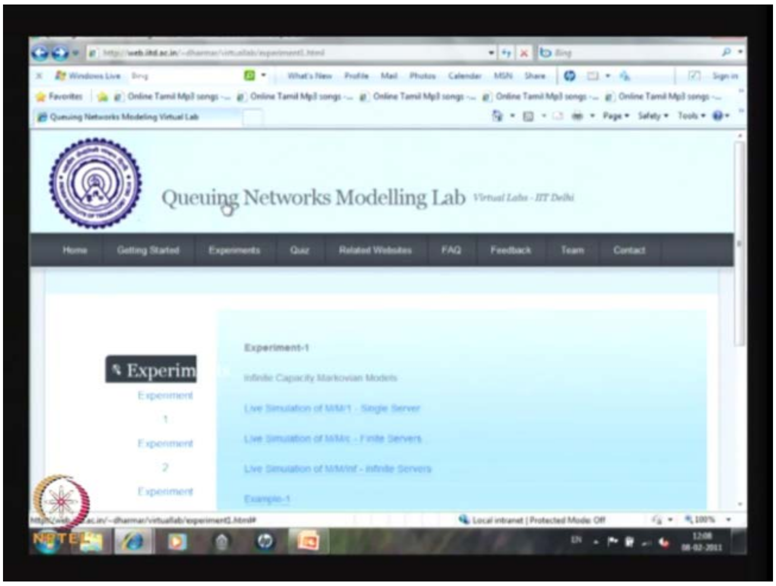
$$q_n(t) = \mu \beta^{n-i}(1-\delta_{0i})\left[I_{n-i}(\alpha t) - I_{n+i}(\alpha t)\right] + \lambda \beta^{n-i-1}\left[I_{n+i+1}(\alpha t) - I_{n-i-1}(\alpha t)\right]$$

$\alpha = 2\sqrt{\lambda\mu}$; $\beta = \sqrt{\frac{\lambda}{\mu}}$; $I_n(t)$: modified Bessel function.

So, use the some identity of modified Bessel function, one can get the solutions pi n of t in terms of pi naught where pi naught we can get it in terms of q 1 where all the q n’s satisfies this equation that is in terms of the modified Bessel function. So, one can see the complete solution in this paper. But here I am giving a very simple approach of getting the time dependent solution for the M M 1 queue by changing this system of differential equation into 1 p d e with the initial condition and solve the p d e. And obtaining the pi n’s and pi naught in terms of modified Bessel function.

Before I go to the summary, let me give the simulation of the M M 1 queue. So, this is the queuing network modeling lab. So, from in this queuing network modeling lab one can simulate the queuing network models. So, for in this I am going to explain how to stimulate the M M 1 queue. And the first experiment that is nothing but live simulation of a M M 1 queue single server as well as you can stimulate multi server queueing model. And you can go for the infinite server model also. So, here I am simulating the M M 1 queueing model.

(Refer Slide Time: 51:36)



Queuing Networks Modelling Lab Virtual Labs - IIT Delhi

Experim

Experiment 1

Experiment 2

Experiment

Experiment-1

Infinite Capacity Markovian Models

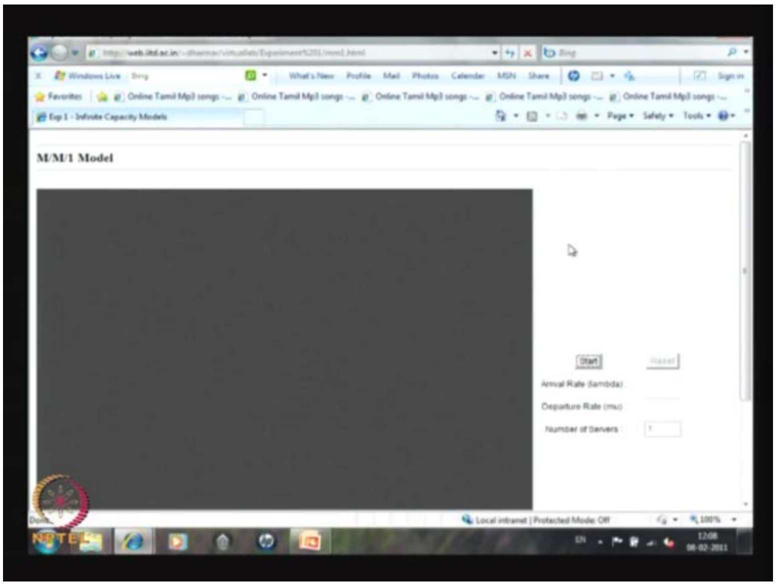
Live Simulation of M/M/1 - Single Server

Live Simulation of M/M/c - Finite Servers

Live Simulation of M/M/inf - Infinite Servers

Example-1

(Refer Slide Time: 52:05)



M/M/1 Model

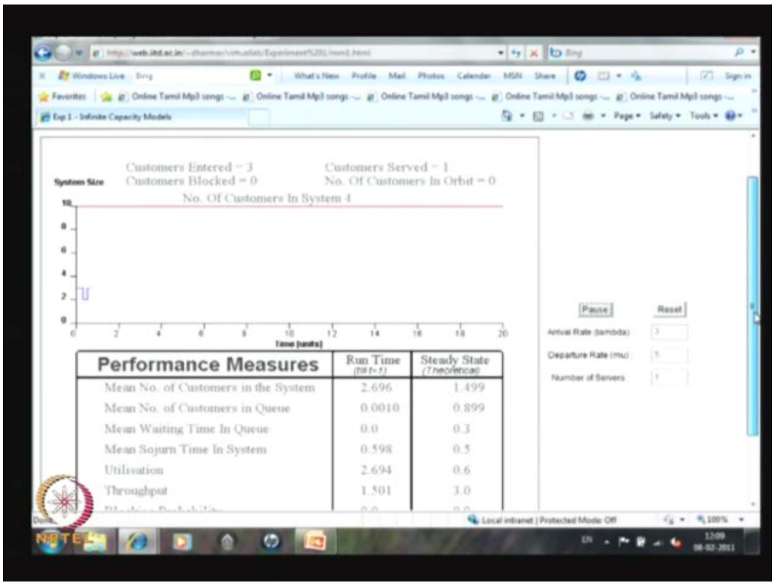
Arrival Rate (lambda) :

Departure Rate (mu) :

Number of Servers : 1

So, to stimulate the M M 1 queueing model, you need the information about the inter arrival time that is exponential distribution you need a parameter lambda the value of lambda as well as you need a value of mu. That mu is nothing but the service rate. So, suppose you supply the arrival rate. Suppose the arrival rate is a 2 and the departure rate is 5 and the number of servers is it is M M 1 queue. Therefore, it is already 1 is placed so number of servers. So, you can start.

(Refer Slide Time: 52:44)



Customers Entered = 3
Customers Blocked = 0
Customers Served = 1
No. Of Customers In Orbit = 0
No. Of Customers In System 4

Performance Measures	Run Time (till t=1)	Steady State (Theoretical)
Mean No. of Customers in the System	2.696	1.499
Mean No. of Customers in Queue	0.0010	0.899
Mean Waiting Time In Queue	0.0	0.3
Mean Sojurn Time In System	0.598	0.5
Utilisation	2.694	0.6
Throughput	1.501	3.0
Blocking Probability	0.0	0.0

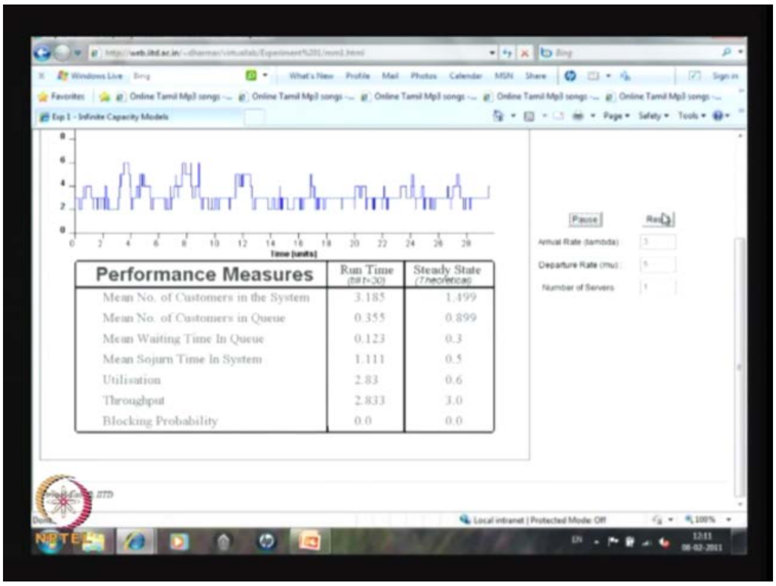
So, this is a way the system increases. So, this is the actual stimulation goes with the, it is time x axis, and y is the number of customers in the system. And here the information is how many customers entered till this time that is 15 customers entered.

And nobody is blocked because it is M M 1 queueing system. Therefore, all the customer who are entering it will be queued. And how many customers are served during this time? And a number of customers in the orbit; this is nothing to do with the M M 1 queue; this is for the retrieving queues. And now, how many customers are in the system at this time. And here this table gives the performance measures the one we have calculated the average number of customers in the system e of r.

And the average number of customers in the queue e of this is mean number of customers in the system that is e of n. The mean number of customers in the queue e of q, mean waiting time in the queue. That is a mean waiting time that is e w, mean sojurn time in the system, sojourn time, spending time or response time all are the same. The mean sojourn time in the system is nothing but e of r. So, this is nothing but the e of r; this is nothing but e of w; this is nothing but e of q and this is nothing but the e of n.

And utilization is nothing, but what is the probability that the. So, here I am giving the run time, what is the average value till this time? And what is the result is going to be in a longer run in a steady state. And the blocking probability is here 0 because the system is infinite capacity model. Therefore, there is no one blocked therefore, the blocking probability is 0.

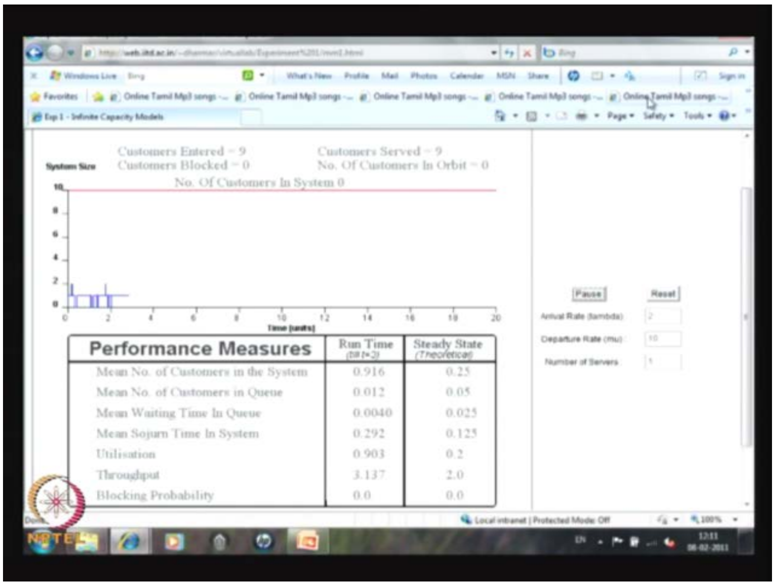
(Refer Slide Time: 55:02)



Performance Measures	Run Time (till t=30)	Steady State (Theoretical)
Mean No. of Customers in the System	3.185	1.499
Mean No. of Customers in Queue	0.355	0.899
Mean Waiting Time In Queue	0.123	0.3
Mean Sojurn Time In System	1.111	0.5
Utilisation	2.83	0.6
Throughput	2.833	3.0
Blocking Probability	0.0	0.0

So, this is the way we can reset, and we can give some other values. And you can start again, and you get the another simulation also.

(Refer Slide Time: 55:26)

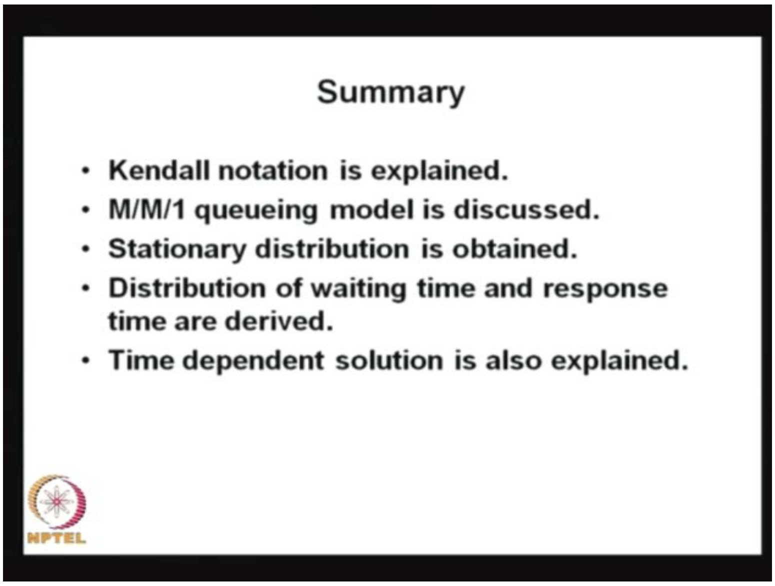


Customers Entered = 9
Customers Blocked = 0
Customers Served = 9
No. Of Customers In Orbit = 0
No. Of Customers In System 0

Performance Measures	Run Time (till t=2)	Steady State (Theoretical)
Mean No. of Customers in the System	0.916	0.25
Mean No. of Customers in Queue	0.012	0.05
Mean Waiting Time In Queue	0.0040	0.025
Mean Sojurn Time In System	0.292	0.125
Utilisation	0.903	0.2
Throughput	3.137	2.0
Blocking Probability	0.0	0.0

And initially it gives the fixed steady state results in the steady state theoretical result. And the run time is nothing but, what is the result for the over the time? With this, let me complete the stimulation.

(Refer Slide Time: 55:37)



Summary

- Kendall notation is explained.
- M/M/1 queueing model is discussed.
- Stationary distribution is obtained.
- Distribution of waiting time and response time are derived.
- Time dependent solution is also explained.

So, in the summary we have started with the Kendall notation. And M M 1 queue is discussed stationary distribution, waiting time distribution, response time distribution is discussed for the M M 1 queue and also the time dependent solution. And I have given the simulation of M M 1 queue also. These are all the reference books.