

Stochastic Processes

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

Dr S Dharmaraja
Department of Mathematics
Indian Institute of Technology Delhi
New Delhi, India
<http://web.iitd.ac.in/~dharmar>



Video Course on
Stochastic Processes -1

By

Dr. S Dharmaraja
Department of Mathematics, IIT Delhi

Module : Continuous-time Markov Chain

Lecture#4
M/M/1 Queueing Model

Contents

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



Stochastic process. This is a module 5 lecture four. MM1 queuing model. In this talk I'm going to discuss the queuing models. So for that I'm going to give the introduction to the queuing models. Then I'm going to discuss the Kendall notation. Then followed by that the simplest a queuing model MM1 queuing model will be discussed. And this is going to be the applications of continuous-time Markov chain in queuing models. So in this lecture I am going to discuss only the simplest queuing model MM1 queues.

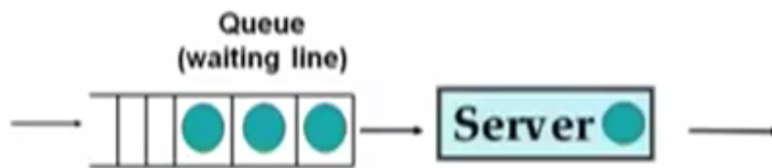
Queueing Systems

- 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



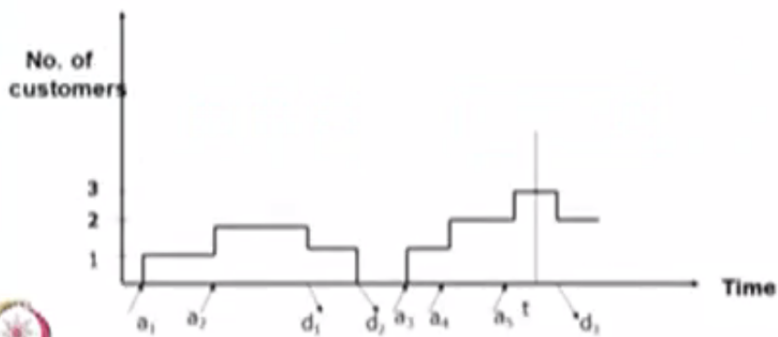
So how one can define the queuing system? You can see many examples in which whenever you go to the supermarket to get some items or you see the railway station counters or you can see the computers service centers many PCs are there and printers and so on so how the queuing system is created and also 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.

Pictorial Representation

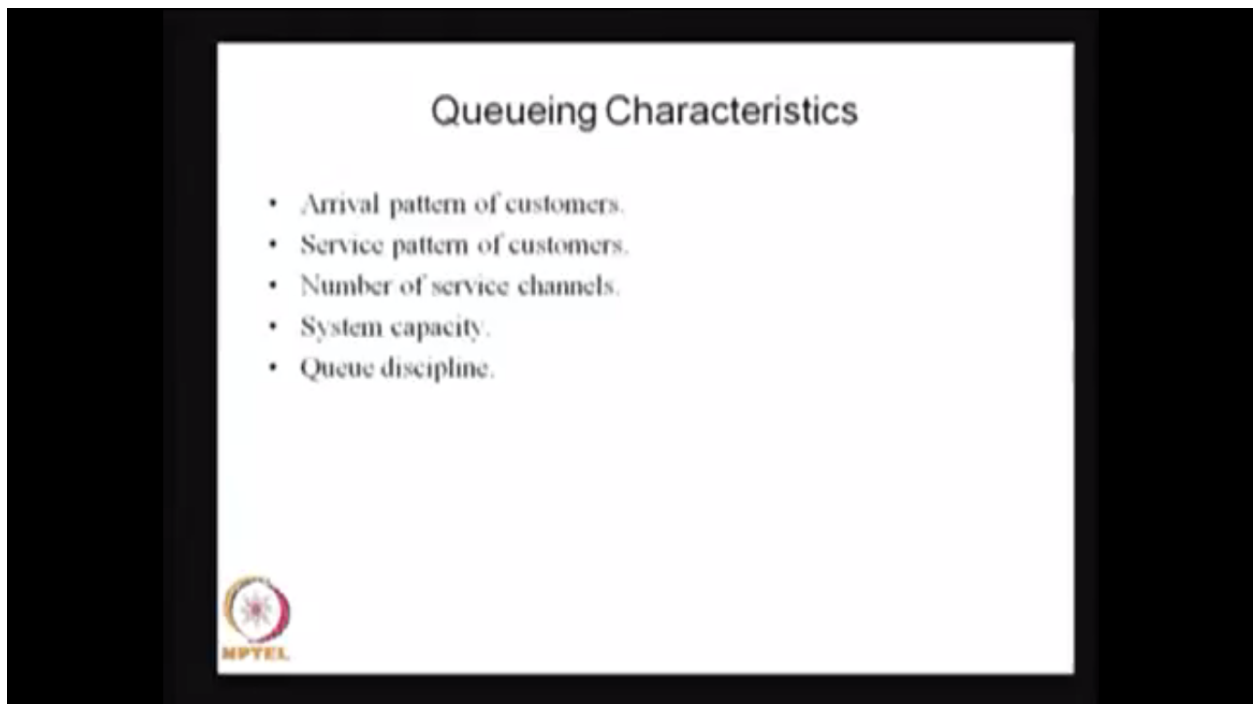


We can give the queuing system we can represent the queuing system in a pictorial form. Some customers are coming into the system and waiting for their service. Once the service is over then they departure from the system. So this is the way one can visualize the queuing system in a pictorial form.

Diagrammatic Representation



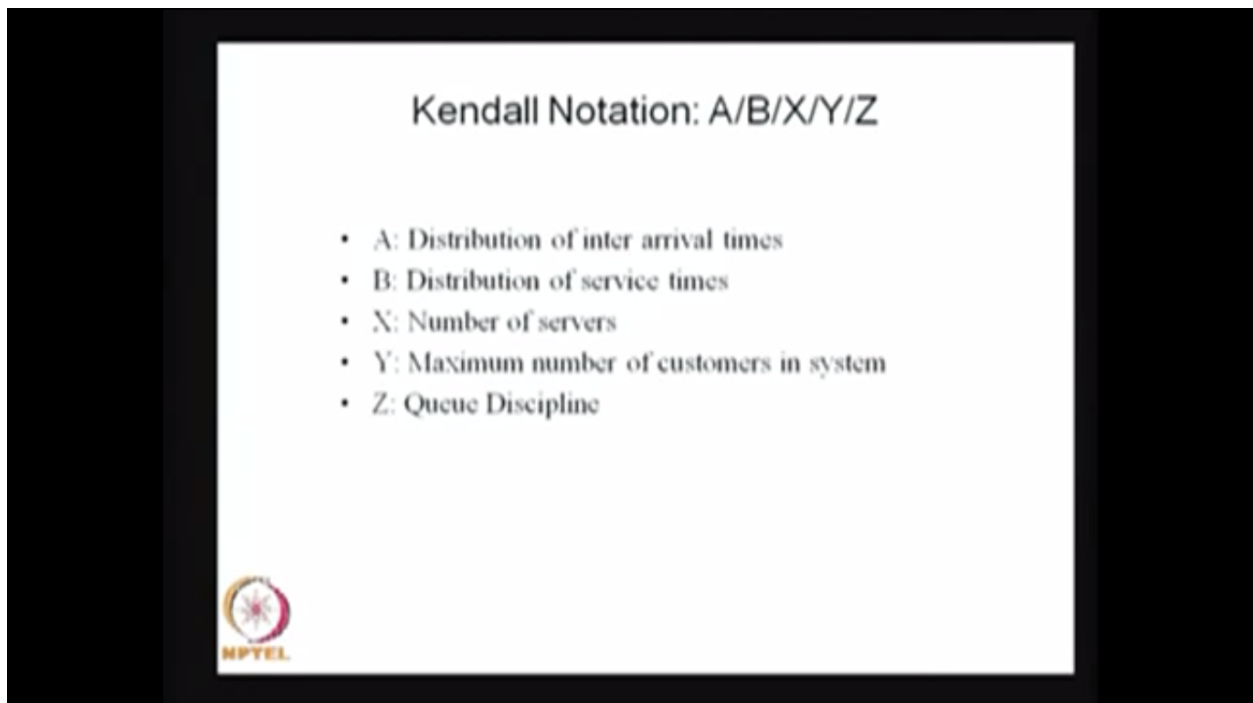
This is a diagrammatic representation and the X-axis is the time and Y-axis is the number of customers in the system. Suppose at time A1 the first customer enter into the system then the number of customers in the system is incremented by 1. The customer who entered the system is getting the service. During his service time the next customer enter the system that with the time point A2. 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 D1 the time point in which the first customer service is over. Now the number of customers in the system is 1. The time point t suffix 2 the second customer service also gets over. Now the number of customers in the system is 0. The third customer enter at the time point A3 so during this interval the system was empty. So like that the system is a keep increasing whenever one customer enter into the system and decreasing by going 1 whenever the service is completed. So this is the diagrammatic representation of any queuing system. Here I made the assumption it's a very simplest one only one customer entering into the system and only one customer is getting served and leave the system and so on. So this is the simple way of a simple diagrammatic representation of the queuing system.



So to define the queuing system you need a few important characteristics. Using that one can easily frame the queuing 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 a deterministic. If the customers are entering into the system with the inter-arrival time that is some random variable then we should know what is a distribution of inter-arrival time. So this information is needed to define the queuing system, the arrival pattern that includes whether it is a deterministic or probabilistic. If it is a deterministic then what is the inter arrival time, that constant thing. If it is a probabilistic then what is a distribution and so on. Similarly after the customer is 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 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 a distribution of the service type.

Then the third important information or the characteristic is a 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 a countably infinite numbers. So according to that the queuing system may be read. So the third information is number of servers in the system. The fourth information that is a 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 wait in the system to get the service. Therefore, the system capacity is also important characteristic. The fourth one queuing discipline. When the customer enter into the system whether they are getting served 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 queuing discipline also important to know how the queuing system is at any time to know that dynamics of number of customers in the system and you should know how the queuing discipline is taken care. Similarly the service discipline also; how the services also takes place during the picking the customers for the service.



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 these are all the minimum important information to characterize the queuing system. One is arrival pattern. Second is a service pattern and the third is a number of servers. The fourth is a capacity of the system and the service discipline or queuing discipline. So based on that the Kendall made a notation and that notation is called the Kendall notation. The Kendall notation consists of A, that letter A/B/X/Y/Z. So the possible values you 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 queuing system. A denotes the arrival pattern information. Here the A denotes the distribution of inter-arrival time. The letter corresponding to the A. The second one 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 that 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 A is for the letter whatever the letter you are going to assign for 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 fourth one what is a capacity of the system. The fifth one what is the queuing 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 these letters.

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. 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 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 the A or B respectively. Similarly if you use the letter E suffix K that means it is a Erlang distribution of type K or you can say Erlang distribution of a stage K that

can be 1, 2 and so on that means in the inter-arrival time is Erlang distributed with the stage K if you place it in the first letter. Similarly H suffix K means a hyper exponential distribution of a type K. whenever you have a inter-arrival time is a 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's a known distribution the only thing is it is other than exponential distribution. So either you can use the letter M, D, EK, HK, or G so G can be other than M itself in the usual or in general form. It's a 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 capacity of the system and the fifth one is the queuing discipline. The default discipline is a first-come first-served. Therefore no need to write the fifth information. And the sixth information is also there what is a population of the customers who are entering into the system. The default the population is infinite that means from a infinite source the customers are entering into the system. That is the sixth information. As long as we won't write as long as the system in which the population is infinite as well as the queuing discipline is first-come first-serve then we won't write. So we write only the first four information; that is a inter-arrival time, distribution, second one is a service time distribution. The third one is number of servers. And the fourth 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 or equal to one that's a multi server system and fourth letter K means a capacity of the system. Suppose we didn't write the forth information here that means it is an 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 we can use a letter general distribution. So whenever G comes in the first place or the second place then we use non – then we say it's a non-exponent non-Markovian queues and if the fourth letter is missing that means it is a infinite capacity system.