Introduction to Queueing Theory

Prof. N. Selvaraju
Department of Mathematics
Indian Institute of Technology Guwahati, India

Lecture - 01
Queueing Systems, System Performance Measures

Hello everyone, welcome to this course. This is the first lecture effectively for this course. As I said in the introductory video, this phenomenon, which we call queues or waiting lines, is prevalent everywhere in our day-to-day life. So, do you like to wait in line for whatever be the reason or whatever be the service that you need? Whether the waiting is physical as you at happens in say in a bank or an airport or traffic jams in the traffic intersections or it is virtual when you know packets are being sent over a communication network or data network or whatever the call centers that then you are calling to get some service. So, wherever there is a system essentially where you need to get some service from that system and whenever you cannot get the service from the system at the point when you want to get it done, you need to wait. So, that is a phenomenon, and you see like these things have been there for a long time. And you design some in such a way that these phenomena can be tackled in some way. So, you expect that in a modern-day newer and newer systems, these phenomena to go away; obviously, it is not going to go away. You know, for whatever reason that you can understand. And you would see a more and more complex nature in the newer systems; whenever you analyze the system, you want a certain improvement. So, every one of us has experience in waiting for some basic service, and in modern-day, more urbanized systems, congestions do happen, and waiting does happen right. Say, for example, if you take on the road, say this kind of thing traffic jams or toll booths. Earlier also traffic jams were there, roads have been widened still traffic jams are there. Toll booths, whether there are enough toll booths, I mean to cater to the commuters who are crossing these toll booths or you know you look at this call-center nowadays all service you get with respect to any of these things you do with call centers. So, in many of these situations, when you make a call, sometimes you may get connected in 5 minutes, you have to wait for some time, sometimes you may have to wait for 20 minutes; it happens. And when you go to supermarkets and restaurants. If you go to a big mall and then there is a big departmental store, and then you want to buy certain items, you put them in your cart and come and wait in line to pay and get out of the supermarket. Restaurants, airports as you see nowadays how things have, I mean exponentially increased the air congestion level both in terms of passengers as well as in terms of the airplanes right and goods transport etcetera, wherever whatever you see. So, you will see that these phenomena of a queue are there, and it was there, it is there, and it will be there, but the important part here is that nobody likes to wait, either the service provider or the service receiver. As a customer or service receiver, you do not want to wait as

soon as you go; you want to get whatever you want to get from that system gets done immediately.

A service provider would like to do, but why wait? Because it is more demand for service than the available facility for service. If there is an airport and a runway, if there is a single runway, how many flights can land and take-off is limited. Obviously, if there are more flights, you cannot do that. So, you need more runways in that particular case; that is the reason I mean. So, in general, you can think that you know more demand for service than the available facility for service. The reasons could be many. Either it could be a shortage of servers, as we just said in runways. Now you can remove this problem of the shortage of servers if you invest in cost; if you invest in capital expenditure, if you do, you will be able to eliminate it by building another runway right. But whether it is economically feasible to do that or there is a requirement, you have to analyze it. Or so, there could be a space limit right in any case; there might be possible that you know you can increase the number of servers in that particular case. So, there has always been, and there will be, and there is a tradeoff between these different conflicting objects from the two different angles. One is from the customer, which we call the service receiver, and the other from the server is the service provider. So, the solution is that you want to design an efficient and best possible system by investing the capital amount. Then if you want to do that now, first you want to study the system; you cannot simply go and invest as you like without any economic analysis done, then you will be not running the system in the best possible way or in an efficient way. So, what does one has to do? One has to first study the existing system's performance if there is one, or one can simulate if one wants and design an improved system and then invest capital in implementing it. So, you need to get answers to certain questions regarding the system you want to design, and then you invest. After doing the analysis, you have to invest in improving or invest capital in implementing the plan. So, to study the performance and to know how many services be made available, one needs to know answers to questions like what is the average time spent waiting in line, how long are the lines on average, how many customers wait more than two minutes, how many customers are turned away, how many servers are needed to achieve a target quality of service, how fast must the servers work and so on. Here, you remember customer here, we mean a generic term it could be me, it could be a physical human being, or it could be airplanes waiting to take off or land, or it could be the message that you sent, and it is in a digital way it is being communicated through a network which is waiting at various nodes for further transmissions. So, the generic term that we use here is what we call customers. Similarly, the server again could mean anything like a generic term here. So, throughout the course, you will see that customer and server are what we will be using. So, the server here would mean again as a bank counter, suppose when you go to the airstrip or the runway or the number of checkout counters. So, the checkout counter is the server, and so on.

This is the generic term that we will use. So, one can relate to the situation under consideration; accordingly, you can do that. So, what queuing theory attempts are to answer such questions through detailed mathematical analysis. So, for example, if you have a production system or a manufacturing system, as we see. Then, you want to know the production lead time of an order when an order is received. What is the reduction in lead time if we increase the number

of machines? Now, what do we do, whether you know we need to prioritize some production over some other components' production? So, if you are going to make a big supermarket and you want to build a parking lot now, how large how big the parking lot should be. So, if it is a city and it is the center of the city, how much will you have to invest in making a parking space for a supermarket. So, whether how large should it be and how efficiently you can decide, that is all. You know like this you know if you want call center of some company say insurance thing right you were making a call now you want to see like how many call center employees or the technician you need what would be their expertise. So, all of them are unique in their own way, or you know their skill sets are the same or different, and how many you need in each type, so you need answers. So, you need to understand first to make an analysis and then decide how many of this. To do all these things like as we see that this is the questions that we raised here. So, these things will become specific from system to system; depending upon the system, the questions will also become specific, and it need not be along with these questions. There may be some other questions that might be interesting when trying to attempt. So, as they say, "QUEUE" is one of the most redundant words in English as 80 percent of its letters are unnecessary because the first letter pronunciation if you pronounce it, you get the complete pronunciation of this word, but queuing theory is not. And also, queuing theory, if you look at it, is one of the words with five consecutive vowels; there are so many things that you can talk about. I mean, about the other side. So, what do we mean by queue? It is a service facility; a system queuing system is a service facility. The queue, like in our words queue, means that the line that forms in the service facilities what we might call, but the whole system also many a time is referred to as simply as queue it might from the context, and it will be very clear whenever you see it. So, a queue or queuing system might be used interchangeably. So, the queuing system is a system in which customers arrive and demand some service from the system, which is for a certain duration before they leave the system, and there could be many different things that can happen when such phenomena can happen.

So, a system can thus be described by the specification of the arrival stream and the system demand for every user, and the service mechanism. So, the former describes the input to the queue, while the latter represents the functioning of the inner mechanism. Because lines form from arrival and arrival here, the customers who arrive and the service process are typically random. So, it is very rare to see that there are exactly two arrivals happening every minute, and there are exactly four people getting serviced every minute. So, these are typically varying whether you take it at a post office or a bank counter or at the airport, or in anything. So, the queuing theory, these are all typically random, the inter-event times in this particular case since they are all random. So, this queuing theory relies on the mathematical theory of stochastic process for its analysis. So, let us talk a little bit about the history of these and the origin and applications of this queuing theory. So, in fact, the first real application of queuing theory that laid the solid foundation for the development of this whole field was in the telephone network in the early 20th century. At the beginning of the century, even you would have seen in a little bit older times that you know how phone calls were being made. From here, you will make, or if you want to call to some other place, you will first make a call which will go to an operator who will then contact through other side, and then he will establish the call. So, this is the

kind of thing. So that was the scenario in earlier days when telephone calls were made. So, a crucial performance measure of such a system is the probability that a person who wants to get a connection for a call finds all operators busy and that, thus, they cannot be served. So this is called the loss probability of a system. So, for a modern-day call center where questions are answered instead of cables connected, the service time represents the call duration between the user and the operator. In the earlier time, it was the waiting that meant the service time cases in this case. So this was the case, and Erlang, a Danish mathematician, statistician, and engineer, was the one who studied this problem and published his work in 1909. So, you can say that the foundation for this theory was laid on that year with that work, and he, of course, did much work on that further, and that laid the foundation for the queuing theory. So, he was the one who first looked at this from the analytical perspective, and how the analytical methods can be used to analyze the systems in that sense was Erlang. So, that is his main contribution and in his honour only in a way like in the forties that telegram traffic measurement unit is being called as Erlangs in his honour.

So, that was he laid the foundation and later on, like people, I mean it was in 1909, 1917 is another major contribution. So, he observed whatever he observed; he put it that what are the two types of models; once we describe the model, we will let you know what models he basically observed as happening in those situations. So, that was the thing that he did, and then, later on, many other people worked on this, and in the 1950s, Kendall looked at the whole queue queuing theory from the perspective of a stochastic process. So that you know that whole theory can be applied here. And the further developments took place when for example, this ARPANET project which is laid the foundation for this the whole worldwide network internet and stuff like that. They realized the queuing theory techniques developed by Erlang and others, and they use this to show that you know how the system was feasible and they established connection is one of those involved was this Kleinrock and from his computer in UCLA to another place is what the first time the packets being transmitted. Anyway, all those things you can find out by yourself. So that is another major breakthrough, and with the advent of computers in the earlier time, whatever the queuing theory which was analytically being made. Now, this advancement in technology can handle more efficient and complex models; that is how it has come up. So, queuing theory is an intricate yet highly practical field of mathematical study. This has many variable applications in fields like computer systems, networks internet, and so on. Internet communication systems networks, whether it is mobiles or call centers, manufacturing service systems, healthcare, airport, theme park, supermarket, inventory management, and so on like, there are plenty of places where you know you will find these applications and what one has to remember is that; most real problems do not correspond exactly to a mathematical model and increasing attention is being paid to complex computational analysis approximate solution simulation and sensitivity analysis. How do we describe, in general, our system, how a queuing system how we depict in a way.

Let us look at a simple queuing system where customers arrive. So, that is what we call an arrival here. Customers arrive, enter the queue, and if there is no one in the queue, they directly go to the service, get serviced, and depart. If there is someone in the server at the time

of their arrival, they wait; as soon as they arrive, they wait in this queue, and as soon as the first opportunity comes, they will get into the server, they will be picked in some way. We are not discussing that right now. They will get into the service, and after service, they will depart. So, this is what the service facility that they have. There are arrivals and departures, and this is the service facility. It is a very simple depiction; this would be the basic model situation that we will be handling in the beginning. Now, as you see, if you add, these are not the only features you normally observe in your day-to-day life. So, there could be arrivals, right, and some arrivals could not enter this service facility for some reason or other reason. So, they are said to be blocked. So they have to leave the system; they will not be able to enter into the system. So, you are going to the hospital, and you want to get admission, and if there is no availability of any bed or anything, then you have to leave. So, you are said to be blocked. Now, once you are inside the service facility, so you want some service again, as usual; if there is one server available, you may be allowed to go to get service; otherwise, you know you will wait in the queue. Now, while waiting in the queue, you may lose your patience. You know you thought that I waited enough and it is not worth waiting anymore. So, you will leave. So, that is, we call impatient customers. And this is the common queue; there is a single queue common queue to be rooted in all these servers. So, this is one server; this is another. There is a mechanism by which you might root to any of the servers, and once you get served, you may leave. Departures happen; these are, again, typically like few features. The more features you add, the more different the model becomes and the more complex the model becomes and hence the more the analysis. However, to understand to incorporate anything further, one needs first to understand the simpler systems and the kind of analysis you do, which is precisely what we will do. We are not we may not be considering very complex systems with many different features, but you should be able to procure yourself with the knowledge that you might take from here.

We might do some of these features and try to guide you like how one can, but we will take very simpler basic systems and try to analyze how one can analyze mathematically and get it. Now how one by incorporation of newer features how one can. So, every system poses a problem meaning that you may find that you know there may be 1000 models available, but none of them fits into this real situation; then what do you have to do? We have to add that feature and build a new model, and that is how the theory develops, and that is how the phenomena happen in this kind of situation. Now we said that our objective is to analyze the system to design. So, how do we know the effectiveness of a queuing system? So, what do you know we will want to do from which we want to gain knowledge about how effective the queuing system is, whether already existing one or you want to design one with certain features? So, now with that features, since you are anyway abstracting mathematically. So, you can build a mathematical model of that and analyze the system. So, basically, there are three types of system responses: how the system will respond to whatever behavior that will happen that one can analyze. So, one is some measure of the customers' waiting time. So, there has been some measure that you want to connect with waiting time, how long you know a customer is waiting, etc. The other is that some measure of the number of customers that may accumulate in the system or in the queue how many are there; that is second. The third quantity is about

the server, like the idleness, idle time, or whatever is related to the servers. Now since the arrival times, the demand requirement, the service capacity, the size of the waiting room, and many more features may be random variables. But typically, even if you keep some of these as non-random quantities, it so happens that arrival is random and service demand, the service requirement is random. Even if you keep any of the other things constant, one can keep those as a random variable. So, depending upon that, even if it is one of them is random, then it induces randomness to the whole system, and the whole system study requires the study from the randomness point of view and performance measures that is what you know we are looking for like one measure connected with each of these three types, they also then become random variable.

So, the moment they become random variables, then you want to know about the random variable, then you want to know about their probability distributions. Many a time again, the probability distributions may be a little difficult to get. So, in that case, what you do is that you try to get certain expected values. Normally, that is what you do with respect to your random variable if you do not know the complete distribution. If you know the distribution, you can answer any question regarding that random variable. But if you do not know, then at least try to get some measures in terms of certain moments and so on, which is what is called expected values are to be determined. Now with respect to the first type of measure, which is the customer's waiting time. There are two types of waiting times; remember that these are all the things that we will consider throughout the course. So, it may be prudent to pay attention to these two different distinctions that we are making because then, later on, it is important that you understand the distinction here. One is the time a customer spends in the queue, and this is called queuing time or waiting time in the queue. So, this is waiting time in queue or waiting time in the queue. This is what a customer who spends time spends in the queue. The other is the time a customer spends in the system, which is effectively the queue, as well as in-service, which is called sojourn time or waiting time in the system—waiting time in the system. Waiting time in queue, waiting time in the system; queuing time, sojourn time. Sometimes the second one is also called system time. Now, depending upon the system, one or the other or both may be of interest. Say, for example, if you are going to an amusement park, where you know you wait outside, that is your queuing time, and once you enter inside, you are getting service, meaning that you are getting to enjoy this park. Or you know you are going to some geological park where a Jeep safari is there. So, you wait outside to get the tickets, and once you get the ticket, you get into the Jeep, then he will take you jungle safari kind of thing. So, then that is the service type. So, in these situations, the time you get service, which is the time you spend inside an amusement park or a zoological park, is not of much concern. You are worried about how long you are waiting outside and how long you have to wait to board a Jeep in such situations. So, that kind of thing. So, that is where queuing time is important, like in the situations; this is the quantity of interest. Whereas on the other hand, there is waiting time in the system. Suppose you have a manufacturing unit, and there the machines are under repair. So, as soon as it has failed, the repair might start, or you know the repairmen are not available. So, it is waiting for it to get repaired, and then some repairmen work on it to get it back to work. So, here it is not just the queuing time that is relevant; the queuing time plus the service time put together. So,

this sojourn time or waiting time in the system is relevant. So, that are the two types of waiting times, and their relevance is in the context of the application, either one or both. Similarly, there are two types of customer accumulation measures which is basically the number of customers. One is the number of customers in the queue the other is the number of customers in the system, which is queue plus service. Suppose if you are looking at a hairstyling salon that you know the number of customers in the queue is what is relevant that you want. Again in the machine repairmen scenario, if you look at it, it is the number of customers in the system, which means the number of machines under repair is what is important rather than the number of machines waiting to get repaired and under repair. Of course, that may also be interesting, but the most important one was this total because you are looking at it from the whole system perspective. Similarly, the ideal service measures can include the proportion of the time a server is idle or the time the entire system is empty right there may be many servers. So, there could be a particular server; what is the idle time, right, the idleness is what then proportionate of that. So, if it is 0.5, I mean typically we express in a fraction; that means, 50 percent of the time, he is idle. If it is 0.9, it means he is always busy. So, you get a certain measure out of that. Similarly, the whole thing can be talked about for the whole. So, what is the job of a queuing theorist or a queuing analyst? Generally, he is concerned with either one of these or both—the first determination of the performance measures and designing an optimal system. Of course, this has to be done even if you want to do the second part here someway. So, for the determination of these performance measures, we must determine waiting delays and queue lengths; these are the typical two things, but depending upon the requirement, there may be other parameters other measures that you might be interested in from the system that you will design. That is the first part of any study, which is the first part, and in our course, most of us will be concerned mostly with this, then one can use this to design an optimal system.

So, in that case, what will you do? Say, for example, one might balance customer waiting time against the idle time of the servers according to some cost structure to determine the optimal number of servers that you would have. You can talk about it; there are plenty of studies on this line. So, again to design for all those things like, you need queues information then, or the first performance measures have to be estimated using that then you can balance out by forming and optimization problem, something like that with the associated cost. So, you have to incorporate, for example, in this particular case, you can associate for each unit of time a customer waits there is some cost involved you associate and for additional each server that you are going to employ per unit of time that there is going to be some cost structure. Like you know, obviously, like if you put more servers, customer waiting time will be less, but if you have less servers, then the customer waiting time will be more then you have to trade off you have to find a balance, right. So, we have to find that. So, one tries to do this problem analytically, meaning estimation of parameters, performance measures, or the determination of performance measure and the first and optimal system, but whenever it is not possible, one does use simulation, but we are concerned with the analytical theory. So, in summary, if you look at it, the problem studied in queuing theory may be grouped differently as the study of the stochastic processes that arise in the evolution and the evaluation of the related performance measures. Once you determine

what these processes are, then the performance measures you will try to obtain. And again, when you want to do that, the method of solution, whether you wanted exact solution or the solution in some form of some kind of transforms or it is an algorithmic solution or asymptotic solution or numerical or you get an approximate solution. Depending upon the nature, you know you will often find some of them. Of course, obviously, one would like to have an exact solution; unless the model is very simple, this may be very difficult to obtain, or even if you obtain it may not be really useful in practice. So, one need not spend much on that. So, one can directly do a numerical technique, for example, or an approximation solution which will give you what the insights are. After all, any mathematical model, you have to keep in mind that it is an abstraction of a real-world phenomenon that does not abstract every feature of the real-life feature; the major features are abstracted into the mathematical model. So, in any way, a model is an approximation to the real system, right. So, you are laying or putting one more layer of approximation in that sense. And again, the nature of the solution is the two types that we will encounter: transient and steady-state. Almost the whole course, except in one place where we will hint about the transient solution; the whole course is concerned with the steady-state solution of these models, meaning this is a time-dependent one, irrespective of the time you are looking at it. Whereas, here, what you are doing is a system is in operation for quite some time, and there is some sort of equilibrium exists, and that situation you are studying the system is what the steady-state solution an equilibrium behavior. Then the problem of control and design of queues where you know you will compare the behaviour and performance under various situations. Again, optimization of some specific objective function involves what they are, depending on the objectives you would want to see in that scenario. Now, what you will do while we do in this course deal with only the quantitative aspects of waiting and its determination through mathematical models and how one can arrive at it. There are certain qualitative aspects of these queues. So, basically, if it is quantitatively, you can decide, for example, this road has this much traffic, and then this is the waiting average waiting number of cars or vehicles waiting at a particular point. So, I want to expand, build one more lane and one more lane I can build, and then if we are in a multistory building, there is a lift, and then you see that always people are waiting for the lift. So, try to install one more lift for it. So, this can be done, but there are certain aspects. So, that means you are trying to improve the experience of waiting. If you are waiting for a lift, and if you have another lift alternatively available, you will obviously be happy that you will take the lift and then go to wherever you want to go. So, you do not need to wait longer, sort of, but that is one aspect. But there are certain aspects where the experience or the psychology of waiting can be improved by looking at the qualitative aspects as well. There are certain principles that one can employ. So, let us look at these principles that one would look at when you look at the qualitative aspects or the psychology of waiting and how one can improve. Again, not everything that looks like a queuing problem can be solved by quantitative aspects alone; please do not conclude in that manner; there could be other aspects that we have to look at. So, that you know, that can also be employed along with, if required, along with this. For example, certain principles that are used to look at the qualitative aspects of waits or these kinds of scenarios. For example, unoccupied time feels longer than occupied time. So, if well, when the customer is waiting, if you keep him busy while waiting, he may not feel that much as

opposed to the case where he will not be doing anything and simply waiting, say, for example, if you enter into a restaurant you are immediately handed out a menu card.

So, you will just be crossing through the menu cards or trying to take; otherwise, as you will simply go and wait, you will feel that you are waiting longer than what is expected. But you do just to reduce you know your feeling of this waiting. Pre-process wait feels longer than in-process wait. Again the same restaurant scenario. For example, if the waiter comes and says yes, ok what do you want like you know menu card if he gives, and then he asks what do you want and then ok he will get a glass of water and keep it to you. So, which means that time your service has started effectively. So, there is no wait prior to that. So, that is a way like you deal with that. Anxiety; will make the waiting experience longer. So, you will feel am I in the wrong line or will I miss the flight or whether I am in the correct position if someone comes and tells you this is what is the thing this is your line. So, you will feel a little bit comfortable, ok, fine, that I am in the correct place and the correct. So, I will get my service, and I will get done, or if someone says I am in a hurry to catch my flight and if you are getting a little late; how long will this line take to pass to the security check and so on. So, you will see that you know if some information is provided, then you will feel a little bit comfortable; that is what it means in this particular case. Again uncertain waits. In a physical queue, you can see the number of people ahead of you in the queue. So, you will probably see that only your service is going to come after their service. So, you can estimate how long it goes may not be at the time of joining, but just after that, after spending some time, you can see how long it will take and so on. But when you make a call to a call centre, you do not know how many people are there. So, if there is some information given to you saying that these many customers are in front of you, it will take this much time to get your service, and in the meantime, they play out all their advertisements and so on. So, this is waiting, and they announce; this one explained unexplained waits uncertain unexplained waits. Again for what purpose are you waiting longer? Suppose you know something happened inside, for example, in the airport like something happened inside some small accident because this you are asked to wait more you will feel impatient in a way, why there is a wait. Suppose there is an announcement saying that ok, there is something that happened. That is why there will be a delay in the take-off of this flight as expected. So, then you know you will feel a little bit comfortable. Again, this unfair way is one of the major things you need to look at. Unfair waits are longer than equitable waits. So, normally we expect that you know FCFS First Come First Serve phenomena to happen, but that may not happen always. So, for example, at a bus stop people come at different times there is no queue when the bus comes they just board one after the other jostling with each other. So, in that case, what happens that it is not that first come first serve business happening there or on the other hand if there is a line formed in the bus stop when the bus comes that people who are there in who arrived earlier get an earlier chance to get into the bus if there is a seat available. Then it is a fair wait whenever you will see this unfair wait. Again this unfair wait in certain cases is inevitable. For example, in a hospital, emergency situations like emergency come. Of course, he gets priority, and if the patient has to be treated immediately. So, he comes late, later than you, but he goes to the service first and then gets his service done. There is always competition with this FCFS business, though. But unfair in quote-unquote in a true sense would really make things longer waits. Longer waits are

tolerable for more valuable service. So, you are in a supermarket, and you have a big cart of lots of items and compare with the situation where again you are in the supermarket with only one item. Now, since you have a big cart full of items, waiting for longer for that may be fair.

Whereas if you have only one item and if you find someone in front of you with a cart full of items, you will feel like I should get some priority reason why you would see that this less number of items you form come to this queue sort of. So, these are all you know situations. So, this is all implied in a way if you really see that is the reason for example, for this particular case, in supermarkets you will see for less number of items there are separate counter for it and for others more the scale. Again solo waits feel longer than group waits; I do not need to explain that. So, this is the case that you would see here in your day-to-day life and everything. So, this is all that you are going to do. So, what are we going to look at it? We are not going to look at any of these aspects. I just thought that you would point out that there are other aspects too when you analyze a queuing system. Apart from quantitative, there could be qualitative aspects as well, which you need to take care when you are trying to address the complete problem. But once you made that it is a quantitative one, how to analyze quantitatively is what we are going to see next. So, that we will see, we will continue in the next lecture.
Thank you.