Communication Networks

Prof. Goutam Das
G.S. Sanyal School of Telecommunication
Indian Institute of Technology, Kharagpur

Module - 04
Circuit Switched Networks
Lecture - 15
Introduction to Queuing Theory

So as we have discussed in the last lecture, we will now discuss the evaluation or performance evaluation of a trunk switch ok. For performance evaluation of a trunk switch we need a slightly different mathematical tool, and that is where queueing Theory comes into the picture.

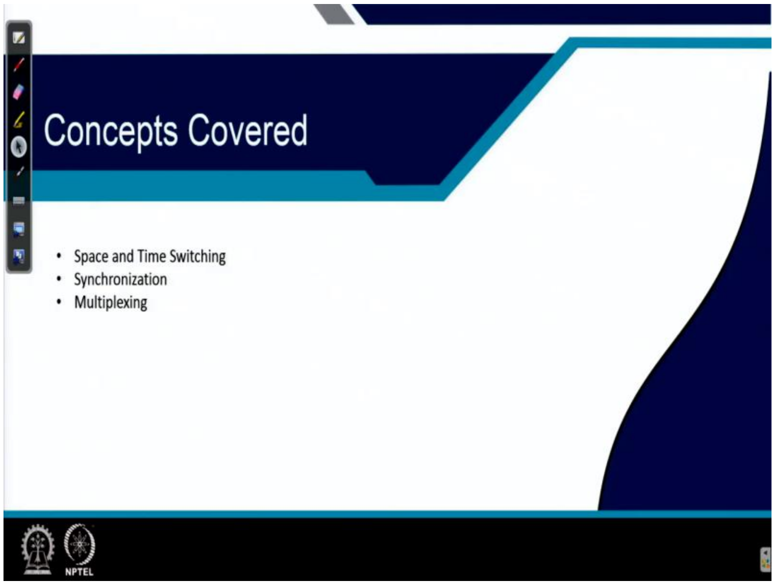
So, what we will do we will try to give a very brief it is not of queueing theory course, of course. But we will give a very brief overview of the queueing theory, as such, especially the continuous-time Markov chain which is something we will be trying to evaluate. And once we do that, we will take that forward, so that will take one or two lectures. And then, we will see how to actually utilize that tool towards analyzing a trunk switch especially the blocking performance of a trunk switch.

So, this is something we will see; we will get a formula for blocking the performance of the trunk switch like we did in the general space switch. That Lee’s formula we have got and from there we were designing switch, over here also the same thing. That how many trunk links should be put that will be designed by this particular formula ok.

So, as such, why we are, you might be asking just for that one formula why we are trying to do this queueing theory. So, queueing theory is having a particular you will see that it has a huge application in networking as such. So, it can be applied again when we will be talking about not the circuit switch network but the data switch network where there are delays we have already discussed.

So, even delay performance queueing theory can be a very handy tool that can be applied for analyzing delays. So, basically for effectively analyzing protocol or effectively analyzing some kind of routing performance. So, for all those cases queueing theory is still a very, very nice tool, and it has a huge application in networking. So, therefore, this course requires some introduction to queueing theory.

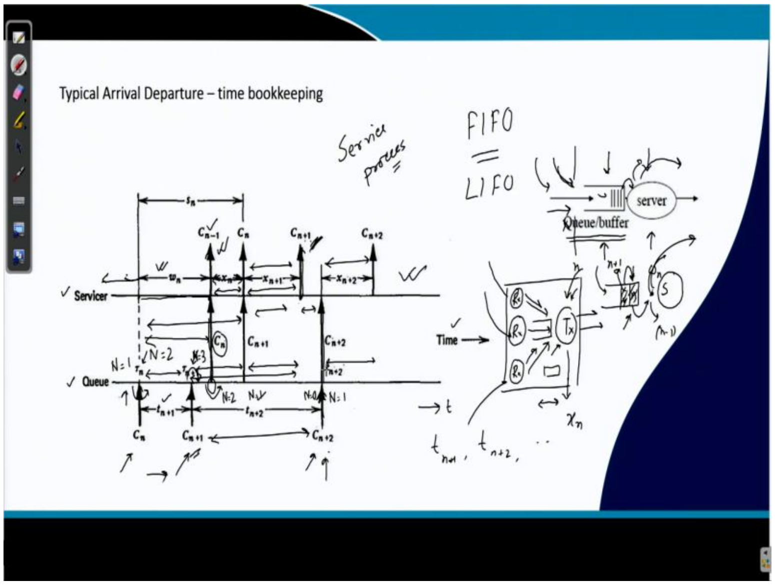
(Refer to Slide Time: 02:28)



Concepts Covered

- Space and Time Switching
- Synchronization
- Multiplexing

(Refer to Slide Time: 02:31)



Typical Arrival Departure – time bookkeeping

So, with that what we will start is we will try to see how a particular queue functions and then we will go into the theory ok. So, let us try to see what generally happens in a queue. So, generally, you might have seen. So, queue we do not have to really discuss a circuit switch network or switch we can see queues everywhere right in our practical scenario, Everywhere there are queues.

So, in queue generally, what happens there are two things one is there is a server. So, basically, if you go to a counter, let us say a ticket booking counter, you go. So, there is a server so; that means the person who is sitting over there who is actually booking your ticket. If you go to the fast food center, whoever is taking orders or supplying you with the food? So, that is a counter, or that is the server we should say?

So, there is a server in the queue, and the server is followed by the most important aspect of the queueing theory; that is the queue. So, followed by the server whenever one customer is being served the other customers are all also there they are anxiously waiting because the guy who is getting served the server is already occupied with him.

Now, all others might not get service, so they generally wait behind the server. This waiting might be termed as a queue; that is what we do generally, we say that it is the queue. That means, wherever you are waiting for some particular application or for some service then you are actually queueing up.

Now, this queue there are. So, basically, if we try to understand queueing always we need to understand these two things; one is the queue itself and the server because the server is a very important integrated part of this queue. So, we have to understand both these things and the interaction between them.

So, as you can see over here also, we have actually put two things over here. So, one is called a queue we might for our communication purpose we might call them a buffer and the other one is the server. The server can be termed as you can see it as a link it is like the trunk links, or we can see that as a router.

So, whatever it is probably the transmitter of a switch. So, something at the last interface where the data is being is going out that is where the server is ok. For data you might be asking ok, what is the service it is taking? So, let us try to understand if you have a transmitter through which you want to transmit data ok.

So, this is let us say this is part of the switch. So, what does Switch do? It actually takes data from some other place. So, basically, it will have a receiver it will receive data from them it might have multiple receivers because it is a switch. So, from multiple directions data might come to those data and might target this transmitter, but what will happen to each of those data if we take that data as some chunk? So, they are being transmitted by the transmitter.

Now, this transmitter is seen as a server over here because it is serving this let us say whatever data chunk you have given that data he is transmitting. So, while transmitting, it takes a finite amount of time which is the service time for our communication server. So, the communication server is nothing, but it is like another server, it is just giving service to a particular entity.

Now over here, entities are not unlike customers in banks or customers at railway counters, or customers in fast food centers they are packets, or they are some chunk of bytes that is coming to the transmitter, which has to be transmitted ok. The service means as much time as it takes to transmit that.

Now, the service how much time will it take to service it? As you can see, it depends on two quantities. One is what is the size of this data chunk because the bigger the size will take more time that is one thing, and the second thing is what is the speed of the associated transmitter if it is serving 1 kilobit per second. That is the transmitter rate; that is how he served.

Then if I give, let us say, 100 bytes then you can easily calculate how much time it will take to serve it. So, that service time will be that much. While he is servicing one packet, that means he is transmitting one packet, and all other packets coming from other sources. They cannot be transmitted because you have a shared transmitter and only they can their packets can be transmitted once he has finished this service like queueing theory.

In all other queues you have seen, that whenever you have finished servicing some in the front of the head of the queue, whoever is there or whoever is there in the server. Once you finish servicing him then, only the other one can enter the same thing happens over here. So, till that time there must be a queue, or we call it a buffer. Because now we are storing the data, so where do we store data in the buffer? So, that is why we call the queues a buffer. So, as I have pointed out, it might be a queue, or for our communication purpose, it might be termed a buffer.

So, in the buffer we store something. So, those data will be stored like a queue, one after another, ok. Now from the queue suppose the server finishes some service after the service time we have discussed about the service time. So, once he has finished the service, then the next customer from the queue will be coming in.

Now, this is where the queue-to-server interaction is very important, so that is where we specify the queue discipline. What do I mean by that queue discipline means? From the queue to the server, how do you select a customer general queue discipline we know that is called FIFO first in, first out. So, whoever is generally in a queue what happens is you put them one after another, and only from the head of the queue do you start servicing them.

So, therefore, whoever has arrived first, he must be getting the preference ok. So, that is called first in first out whoever has come first he must be served first in the queue. So, that generally happens, but not necessarily always queue discipline will be like that. There is queue discipline which is LIFO; that means whoever has come last will be served. Sometimes it might also happen that from among that customers or among the packets in the queue, whoever requires the least amount of time to serve you generally picks them.

If you have read computer architecture or whoever is having a computer science background, you might know processor scheduling, there are so many types of these kinds of schedules which are actually queue disciplines from the queue where jobs are stored how do you serve the job?

So, least job first, sometimes you do the highest amount of time that is taken for a job that you serve first, sometimes you do it completely randomly. So, we just pick among them assign a probability, and randomly pick customers, so that also you can do. So, there are multiple disciplines of the queue that might happen in queueing theory.

So, that is something we will be discussing later, but right now, probably the fairest queue we should say, is the FIFO queue we will be discussing about that queue. But there is no restriction to that there might be a deviation from this particular FIFO queue as well ok. So, once we have understood this particular part now, this is something where we need very highly focused time tracking. That means, what are the events that are happening in this process? Let us try to characterize these events a little bit.

So, what do we mean by the event? A customer arrives at the system, that is an event the customer or maybe packet arrives to a system. It is getting queued over here that is the event. It is entering the first customer entering to the server or the server whoever is being served he is living these are all events ok.

So, these events will be happening at different times. So, we need to actually track these events and try to see what is random in those events. It is a time-varying process, and if there is something random then this will be characterizing a random process. So, queueing theory essentially talks about a random process where there is a randomness we will see where the randomness comes from, and there is an associated time variation of this randomness. So, that is why it is a random process.

So, if we have to make any meaningful sense or essence of queueing theory, then we will have to try to understand this particular random process, and we have to characterize that random process. Let us try to see and appreciate this random process. So, over here, what I have done, we have two entities; one is a queue one is a server. So, I put a timeline this is where the time goes ok. So, time goes in this direction left to right ok. So, now, let us try to see the events.

So, over here what has happened is the C_n is the event; that means this is the time which is marked by τ_n. So, at tau n time a particular customer has entered the nth customer, or I should say I am probably from time 0 I am counting at tau n the nth customer has entered into the buffer or entered into the system ok. Whether he has entered into the buffer or not that we will see later on ok. So, there are cases that he might not enter into the buffer.

But let us try to understand that this is the nth customer if outside the queue some gatekeeper is there he is seeing that one by one customers are coming, these arrivals this is called the arrival process these arrivals are generally random. Because what time a packet will be arriving or what time I will initiate a call that is completely random, nobody can actually, from the system perspective nobody can guarantee, it is coming from outside the system, and it depends on the whims of the customers who are making calls or who are going to some counter.

When you are going to a fast food center that the fast food center does not know, it is dependent on you. So therefore, for him, it is a random thing, and not only that, multiple customers are arriving. So, it all depends on their individual whim whenever they feel like going over there ok.

So this is called the arrival process. So, this C_n demarks the event of the arrival of the nth customers in the system ok. $C_n + 1$ it is the arrival of the n plus 1st customer it arrives at $\tau_n + 1$ this sorry τ_n this arrives at $\tau_n + 1$. In between they have some amount of gap, this gap is between two customers independently arriving at the system ok.

Similarly, C n plus 2 is the arrival of n plus 2 eth customer to the system at tau n plus 2, and this gap is demarked by t n plus 1 n plus 2 right. Now this inter-arrival time inter-arrival time means the time between two successive arrivals, these are also random because this depends on the whim of all those customers.

So, customers $n, n + 1, n + 2$ when they will be arriving that is completely random and they also do not coordinate among themselves; they are completely independent they make their arrival independently. So, therefore, the inter-arrival times are going to be random things ok.

So, these times this t n plus 1, t n plus 2, and so on these times which are counted as accounting for the inter-arrival time between successive arrivals, they are also a random parameter ok; an associated random variable they have and it. Because it varies with time, so it actually makes a random process and that is why the arrival is called the arrival process which is often characterized by a random process ok.

Good, we have now understood that there is something called the arrival process which is a random process. We were saying that in a queue, we need to understand what are the random things that are happening. So, this is one random thing that is not under the control of the system it is coming from outside, it completely depends on the whims of the things that are arriving. If it is a packet, if it is a call, if it is some customers, so according to their whim it comes. So, this is one random process.

After their arrival, so there are arrivals are some events or some instances of something is happening, so these are the events that are happening. Now after the arrival what might happen? So, he arrives over here let us say the nth customers he arrives he goes inside the queue or inside the system then he sees there is a server and there is a queue. There might be two situations that might happen it might happen that there is a customer already getting serviced.

So, whenever you enter a queue you might see there might be two situations; one is already the server is busy, then what do we do? You join the queue. Or you might see that the server is free nobody is waiting in the queue of course, so if the server is free; that means, nobody is waiting in the queue; that means, in the system, nobody is there.

Then you do not join in the queue unnecessarily you directly go and join the server. So therefore, you have two decisions to make or two decision two particular decisions that you can take; one is whether I should join to the queue or should I join to the server, which depends on what is the current state of the particular system.

State means where exactly it is at that particular time instance tau n where the system is; that means how many customers are there in the system. Is there a customer who is getting service? How much time he has already taken the service? How much time is left? Who are the other customers who are waiting? How many of them are there? How much time will they require to service? So, that is the overall information is the state at that particular instance ok.

The state of a system also it is a time-dependent thing, so that is also a random process actually, and it is random. Because at different times because the inputs are random. So therefore, the states also will be random. So, at a particular instant, state means all these descriptions that I have just given ok.

So, with those descriptions, he tries to see the summary of the state, and then from there, he decides, whether should I join a queue or should I join a server. If nobody is there in the system then he definitely joins the server, if there is somebody, then he joins the queue ok. And according to the queue discipline if it is a FIFO queue, then he joins the rear end of the queue ok. So, that is what he does, and he keeps waiting till all the customers before him are being served then only he will join the queue ok.

So, at this instance, he joins the queue let us say, because maybe for this example, we have seen that C n has seen some customer is already there in the service. So, this is the instant I am seeing the server, and this w n demarked the amount of time the means customer who was in service when C n enters he will take to get his service done ok.

So, that means, let us say for packet a packet transmission is going on or let us say the means just before whatever packet was there that packet transmission was going on or maybe something was going on ok. It might be just before then he will be he will be the first customer in the queue if, or he will be the first packet in the buffer. If there are some more packets, then he will be that many after that many numbers of packets ok

So, over here, the example we have taken is actually something like this, you have a server only one packet is being served by the server, and then the queue is rest of the queue is empty, and you will be the first one joining in the buffer ok. This is the nth customer or nth packet that joins the buffer. And the server is serving the packet, which is n minus 1st packet and that has some amount of time left, from there on, which is W_n. So, w n amount of service he is left with at the instance when this customer joins the queue ok.

And then what will happen as time pass by there are multiple things that might happen. So, over here, we have taken an example. So, in that example, we have told that maybe the next customer will arrive before the service is finished before the n minus 1st customers service is finished the second customer sorry $n + 1$ with customer arrives. Then what will happen because the nth customer is over here $n + 1$ customer will join in the second place of the queue.

So, he is still joining he is there waiting in the queue at W_n this customer's service is over. So, he leaves immediately there will be a chain of events simultaneously that will be happening. So, he leaves; that means he departs his departure he is joining the service and he is updating of the queue location. All three things happen simultaneously $n + 1$ customer will be now at the head of the queue, the nth customer will be at the server, and $n - 1$ the customer departs the system.

So, all these things happen over here, ok. So, the nth customer this $C_n$goes to the server and $C_n - 1$ departs over here ok. So, all these things are simultaneously happening and of course, $n - 1$ customer actually changes the queue. After that this is the amount of time this x_n this nth customer takes for his service ok.

So, this x_n time there will be absolutely nobody is arriving because the arrival is the next arrival is scheduled at a much later time. So, there will be absolutely nothing happening. So, he will finish his service he will depart at this instance at that point, the customer this $n + 1$customer who was in the queue will join the server and then he will start his service.

As you have seen in the queue when you go to the queue taking service is it always true that you will be taking same amount of service? Not necessarily. Let us say you are going to a food court and you are ordering. Some will be ordering some ready made things some sweets. So, it is just has to be given. So, it has to be collected in a plate and then has to be given and then the money will be taken from you and that is it.

But you might order something more fancier which has to be heated up and then it has to be given to you. So, basically every service has its, own characteristics and accordingly it will take time. For switching network also or packet switch network also or any kind of circuit switch network, that is also different. Because how much time you will be talking or how what is the length of your packet that completely depends on what kind of packet you are bringing.

So, therefore, we are now getting another random things over here which is called the service process. That means, how much amount of service, I am bringing what is my packet size? What is the kind of service I am bringing? Am I booking a platform ticket? Or am I doing a long distance reservation at the railway counter? It depends on that. So, what service I am bringing accordingly I will be taking service time.

So, therefore, service time is also not the characteristics of the system it is defined by the incoming customer. So, which is coming outside from the system coming into the system, outside from the system. That is another things which is called service and this is also making a process because it is also random the it is depended on customers whim depending on the customers there will be services which should be required.

You do not know exactly which customer will be bringing what kind of service it is up to them and they will be randomly arriving and they will be also randomly bringing different services. Today you might go and book a platform ticket tomorrow you might go and book a long distance reservation or do a long-distance reservation which requires a lot of filling out forms and all those things.

So therefore, this particular service process it is also a characteristic of the input to the system not it is a characteristic of the system. So, as you can now see we are almost on the verge of means characterizing two things of this queuing, one is the arrival process the other one is the service process; both things characterize the whole thing ok.

So, we have to whenever we are talking about these things we need to characterize the arrival process, and also we need to characterize the service process ok. It is a time-dependent random thing most of the time each customer are arriving independently of the other most of the times which is a very fair assumption most of the times each customer or each packets are bringing a random amount of service, and they are independent to each other.

So, if this is the case then we need to really characterize these two things and they are associated randomness ok. So, we go along with this let us finish this discussion. So, basically again this $n + 1$ customer will take $x_n + 1$ amount of time and then he will be departing. Fortunately nobody still has arrived, so at this point what will happen.

Now, you see that situation where the server and queue both become empty because no new customer has come all the customers the $C_n + 1$ that has already this $n + 1$ customer that has that was the last arriving customer that has already departed. So, no new customer has arrived. So, this is the time when server will be free.

At this instance whenever the next customer arrives $C_n + 1$ you can see that it has two events simultaneously. He does not actually join a queue, he directly bypasses the queue and goes and joins the server and then he gets his service at $C_n + 2$ he departs in between if there are. So, it continues, so this is the process.

So, as you can see both the randomness of this inter-arrival time and this service time these two together interact in a complex manner the way I have described with the queue and server. This means this conjugate process of the conjugate system of queue and server they interact with each of them, according to the queueing discipline basically forms this complicated random process, and this is the random process I have to characterize.

You will see how do I characterize this random process let us let us try to understand. First, let us give the state description if my state description one of the state description is how many customers are there in the system, that is one of the descriptions ok. So, that is how do I describe that at this instance. How many customers are there? This guy was getting service probably from a means time before. So, at this instance, only one customer was there in the server nobody was there in the queue.

So, therefore, number of customers if I say n that was actually 1. After the arrival immediately the state changes. So, state goes to there was one customer who is already getting service, and another customer has joined in the queue. So, n has become 2 at this instance another arrival has happened, but the same customer is being served.

So, therefore, n has become 3 at this instance what has happened one customer has departed this n minus 1. So, you are left with two customers, so n has become again 2. Now over here another what has happened another departure has happened. So, over here n has become 1. Now the next departure has happened no arrival, so n has become 0 over here. Another arrival happens n has become 1.

So, as you can see if n is a particular random quantity that is over time getting changed and every event is making a change to its state. You should also very carefully see that its changes by only plus or minus 1 ok. We will discuss about that later that is that is a very important criteria for this, or the kind of things we will be discussing later on, but this is what happens.

This is the state evaluation, and this state will be trying to track down that is the most important part of our job that is one thing we will be doing. Here there are lot of other hidden information if you try to see one of the hidden information is how much time each customer has waited.

Let us see this customer has come over here he has in the queue, how much time he has waited? This much time he has waited in the queue, after that he has gone to the server and this much time he has waited for service. Overall his waiting time was this much the queueing time plus the service time.

Now, if you see this customer how much time he has or this customer let us see this customer how much time he has waited. He has waited as you can see this much amount of time for queueing and this much amount of time for service. So, this is the overall time he has waited over here for getting the whole service done. For this guy no queueing time this is the service time ok.

So, basically for each customer if you have this timeline that event occurrence history then you will be able to track down how much delay each one has got. So, I get the essence of delay over here. So, this is a very nice tool if we can track down this and mathematically analyze this. Probably we will be able to also, characterize the statistical property of this delay which is a very important thing from the design perspective.

So, all in our queueing theory we will be trying to do or track down is how do we actually do this system analysis to get those very important insights of delay or some other parameter we will see later on. So, the delay means queueing delay or overall delay or service delays all those things, or sometimes we will see blocking probability these kinds of things also we will be analyzing. So, we will see how to do those analyses in the next class probably ok.