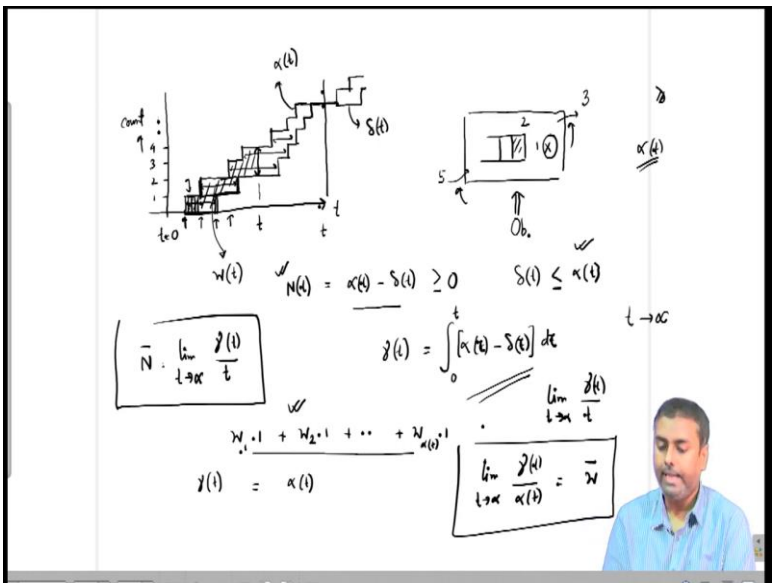
Communication Networks
Prof. Goutam Das
G. S. Sanyal School of Telecommunication
Indian Institute of Technology, Kharagpur

Module - 05
Queuing Theory
Lecture - 21
Little's Theorem cont'd

Ok. So, we will start our means we will continue our discussion rather on Little's formula. So, we have been discussing about Queuing theorem and we have also characterized the input process which is characterized by two random things: arrival and service.

So, that is something we have seen we have seen a typical memoryless arrival and service process. So, those are the things we have already seen. Now, we want to derive a very simple formula which will be very useful that is the average relationship between the number of customers in the system and the average amount of time they will be delayed or they will be waiting, sorry will be spending in the system and the average arrival rate. So, we want to derive the relationship between these two.

(Refer Slide Time: 01:21)



So, let us try to see how we can do that. So, for that we have started discussing a simple process again it is book keeping ok. So, from that bookkeeping will be able to get the inside ok.

So, like we have been doing for this particular part of the course. So, this bookkeeping is a little different. So, what we are trying to do, we are also trying to keep a count of this bookkeeping. So, this is where we will be keeping a count ok.

So, this is just countable. So, it goes on 1, 2, 3, 4 and so on ok. What we are exactly doing over here? We are trying to so basically we had a system if you remember. So, we had a system that had a queue followed by a server and this is the whole system, the customer arrives and they depart ok.

So, somebody who is the observer. So, he observes the system from outside and he tries to see the count of this one and count of this event. So, these two are the most important events from the observer's point of view for the system.

So, he just tries to see who is coming at what time and he is keeping a count starting from t equal to 0 ok. So, this was our scenario and then we will have the arrival count and we will have the departure count.

So, we have started doing that it will be like a monotonically increasing staircase kind of thing we have already discussed that. So, whenever the first arrival happens it increases to 1, then the next arrival happens it goes to 2, then the next arrival happens then it goes to 3 and so on it just keeps on going like this.

And when these increments will be happening that are all initiated by an arrival ok and they are basically random. So, it might be exponentially distributed this inter-arrival. So, accordingly, things will be happening.

So, we are calling this curve to be alpha t, the arrival count curve. Similarly, we can put a departure count curve and we have also discussed that the departure count curve will be always below this, occasionally it might touch the arrival when there is nobody in the system; that means, as many have arrived that many have departed but otherwise it will be below this ok.

So, this will be always true. So, now, let us try to see how we can put the departure curve. So, the departure curve will be similar things whenever there is a departure happening. So, there will be one count increment ok.

So, that is an arrival-departure curve. Occasionally what might happen that the departure will probably let us say it is going like this. So, it might catch up, and then again they might divert something like this.

So, this is happening. We have also discussed that of course, they might catch up, but delta t will always remain less than so if this is the departure curve delta t or departure count curve will always remain less than equal to of course, sometimes it might become equal so, but it will be always either less than or equal to this alpha t this will be always happening.

From there, we have also taken one inference what is the height of this one at any instance? So, at times t the height says it is the difference. So, height just says that it is the difference between alpha t and delta t the naming that to be N t and we can easily see that is nothing but the number of customers at that time inside the system because if 5 customers have arrived and the count says 3 customers have departed 2 must be inside one of them will be in server, one of them might be in queue something like this.

So, always this difference gives me the instantaneous value at that particular instance of time how many customers are there inside. So, therefore, this is the count or number of customers that for the average we want to link to the right.

So, we already from this very simple curve that we can see that we have a relationship with this N t is alpha t minus delta t and this is always as the number of customers has to be either positive or it can be 0. So, because we know this relationship alpha t is always greater than equal to delta t. So, this is always greater than or equal to 0.

So, that is also consistent this is very good. Now, let us also try to see that this is the vertical part. Horizontally, what do I mean by this? Horizontally as we can see this is the arrival instance of the 1st customer and this is the departure instance of the 1st customer.

So, therefore, horizontally whatever we get from one boundary to another boundary of this alpha t and delta t curve that is actually the waiting time of every customer. So, this

is the waiting time for 1st customer, that should be the waiting time for the 2nd customer because he has arrived over here and he has exactly departed over here ok.

So, similarly, this should be the waiting time for the 3rd customer and this should be the waiting time for the 4th customer and so on it just goes on ok. So, basically, we get this with this we get the waiting time ok.

So, now let us try to see how we infer this whole thing. Let us try to do one more thing that is this particular thing alpha t minus delta t that this difference curve if we just integrate it from 0 to t d t of course, will put t tends to infinity. So, we will integrate it for a sufficient amount of time because we want to get the average values.

So, what this gives? If we see the curve this actually gives the shaded region ok because alpha t integration minus delta t integration. So, it is the difference between these two. So, that shaded region gives that portion actually. Let us declare this as some gamma t ok.

Of course this variable maybe I should put a dummy variable because I am putting the limit to t. So, therefore, that is a dummy variable tau we have put and the limit is up to t ok. So, let us call that gamma t. So, we have got some integrated value. Now, let us try to see what is this gamma t. If I see it in the horizontal direction so basically it is all instantaneous value N t it is the integration of that up to time t.

So, therefore, the average of N t is talked about as N bar. What is this? We should take over the time average. So, that should be 1 by so basically we are trying to take the average of this one. So, over time if we average it. So, therefore, that should be gamma t divided by t, or rather what we are trying to do we are integrate that whole thing 1 by t is divided by time and we are taking limit t tends to infinity.

So, if we just do that. So, this should be gamma t divided by t and we will put that limit t tends to infinity. So, this will be my N bar. So, therefore, I can just rewrite this that should limit t tends to infinity gamma t divided by t must be N t.

Now, in this particular integrated area, I can also see what is the height. So, this is actually the aggregation of these box areas. What is the height of each box? That is exactly 1 on the x-axis sorry y-axis we are just given a count. So, therefore, that will be

that height will be 1. And what is the width of that? That is actually the waiting time for each one of them.

So, let us say up to t how many arrivals have happened? So, up to t delta t this alpha t arrival has happened right up to t alpha t gives the count. So, up to you if I take it. So, let us say this is my t up to t how many arrivals so far have happened? Alpha t number of arrivals has happened ok.

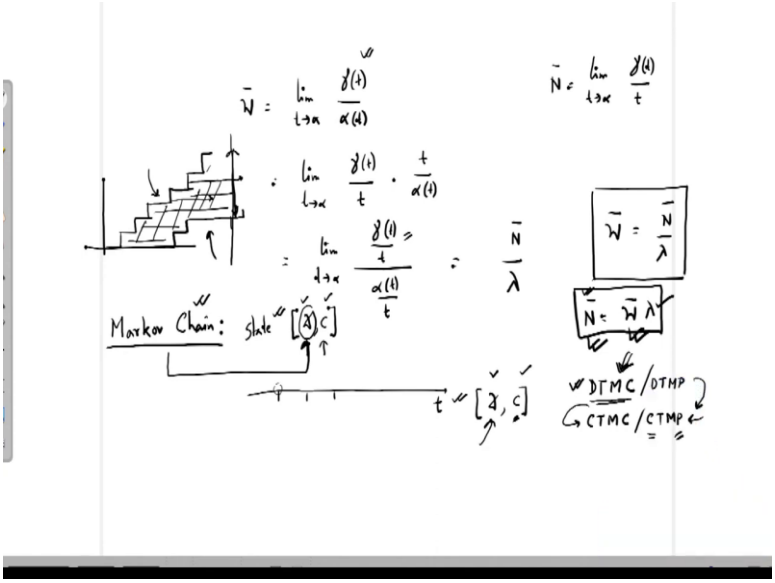
So, basically what I can say is what is now my overall area that should be as many arrivals has happened that many this area has to be accounted for. What is each area? Each area is for the first arrival let us say arrival 1 ok. So, his waiting time is W 1 into 1 plus the next arrival two his waiting time is W 2 into 1 plus so on up to how many will be happening.

So, W alpha t number of into 1. So, basically, this is nothing but this area gamma t I can also write as alpha t number of these things are happening. So, alpha t number of cases into the means length of waiting time or length of system time ok.

So, now if we try to actually divide this means I want to do an average of this. How do I do that? So, I take all these values divided by the number if I divide that I will get the average value. So, therefore, this gamma t which is these things if I now divide by alpha t then I get the average waiting time.

Let us call that W bar, of course, again limit t tends to infinity I will have to take ok. So, these two relationship I have already got. Now, let us see how they are related. This gamma t by alpha t I can write it in a slightly different order.

(Refer Slide Time: 12:27)



So, if I just. So, what we have got? W bar we have got as limit t tends to infinity gamma t divided by alpha t ok. Also, we have got an N bar limit t tends to infinity. So, that was gamma t divided by t, right? So, these two relationship we have got.

Now, this I can write. So, limit t tends to infinity I can write this as gamma t divided by t into t by alpha t or I can write limit t tends to infinity gamma t by t divided by alpha t by t ok.

So, limits if I distribute them, and then this one I can already observe that must be N bar. And this one what is this? In t amount of time, alpha t amount of arrival has happened. So, if I divide by t that should be the average, per unit time how many arrivals are happening? So, that must be lambda which is the famous formula, or Little's formula.

So, the W bar is the N bar divided by lambda or we can say the N bar is nothing but the W bar into lambda. So, the average arrival rate, average system time, and average number of customers in the system are interrelated. As you have seen while deriving this I have never put any other condition on how the queue has to behave is it FIFO, LIFO means random service we have not discussed anything.

So; that means, it is true for any other things ok. So, because we were just seeing how many customers were arriving, and how many customers were departing ok we never considered what was happening inside. So, therefore, it is irrespective of this relationship is true irrespective of whatever you do in the queue, whatever discipline you put in the queue.

Also, we have never talked about how the arrivals are happening, what the distribution they follow or what is the service distribution we have also not taken that thing into account. So, therefore, it is irrespective of that also.

So, what we can see this is a very strong formula that is always true irrespective of what kind of arrival process we take, what kind of service process we take, what kind of queue discipline we take, how many servers are there, what is the arrangement of those queues, how those queues are linked to the server.

So, for all those internal arrangements are it is this formula is oblivious to that; that means, it is true for all those cases. So, it is a very strong formula in queuing theory and this is after Little so it is called Little's formula it is a very useful formula and this actually gives you a general lambda that will be known. So, if you can calculate one of them you immediately get the other one.

So, basically with effort if you can calculate one of the parameters immediately we will be able to link to the other parameter; that means, if we can get an average number of customers immediately we can get the average delay ok.

So, that is the relationship it gives it is a very handy formula and it will be we will see that it is heavily used in queuing theory of course. So, that is the whole thing of course, over here you have to do some observation I was putting t tends to infinity and I was able to do all these things, but just carefully check whenever we were drawing this thing to remember if my delta t is unable to catch up this ok.

So, then alpha t will keep on increasing. So, these differences actually keep on increasing so; that means, the average number of customers waiting in the system will be increasing. What does that mean? If my system is such that the arrival rate and the departure rate there is discrimination. The way things are arriving I am not able to capable of servicing them as fast. So, my queue will build up infinitely.

So, therefore, the system becomes unstable we will when we discuss queuing theory we will talk about these particular things where the queue becomes unstable and where we cannot do any analysis either we cannot do any analysis that is a separate thing, but in a real system also it will become unstable the queue will infinitely be big and the delay will go under means without any control. So, that kind of thing will be happening.

Over that formula is not applicable you know why because up to t we were saying that we are counting the delay. Our assumption was up to t as t goes to infinity they will be close enough. So, that everybody has almost finished their service or most of the customer has finished their service, but if all the customer up to t has unfinished this things.

So, then I cannot count the delay this area under this is not equivalent to that W 1 plus W 2 I can write up to this one let us say W 1 and W 2, but W 3 still have not finished. So, I do not know whether I can take that W3 3 it still has not been finished. So, if most of the cases are this then that formula is not valid.

So, for this formula to be valid we need to have a stabilizable system that is very important; that means, this gap between alpha t and delta t should be close enough. I cannot say exactly if it will be the same, but statistically, we can always say that they are always coming close by occasionally becoming 0 again deviating a little bit and again coming close by. So, that is the pattern you are seeing therefore, for most of the packet you will be able to finish the service and you can account for them.

So, therefore, your calculation of t tends to infinity and will become more and more accurate. So, this average calculation will give you an accurate value. But for that the system has to be stable if the system is not stable now the packets will be probably serviced and it will have an infinite number of packets that are not serviced they are partially serviced.

So, you cannot account for their delay because it still has not finished its service. So, this is something we have to keep in mind whenever we are putting the Little's formula. Also, this is in general in queuing theory we will always have to keep these things in mind ok whenever we are doing that.

So, with this, we have got a basic background ok. So, we have seen the average analysis, we have seen characterize the arrival process, we have characterized the service process. Now, what we will try to do we will go ahead a little bit and then start analyzing it and that is when we will be doing something that was done by Markov it is called the Markov chain.

So, we will do this Markov chain analysis ok. What is a Markov chain? What does it signify? That is something we will discuss, but let us try to understand some of the properties of this Markov chain.

The first thing is in this process what we are talking about is a queuing process where packets are coming they are being served by a particular transmitter or where means some people are coming and there is a queue, there is a server it might be anything it might be fast food center, it might be railway counter, it might be other reservation counter, it might be in airport passport control things whatever it is there are queues everywhere ok.

So, that kind of system if you try to see the description of that system is a random process of course. So, it has a time description. And we have also seen that it can be described by some in time some events are occurring. So, it might be described with respect to those events or we might rather say some state. So, every time something something is happening and then we actually at every time whenever there are events which are occurring we actually try to capture that by some state of the system.

So, we accordingly define the state of the system; the state might be the number of customers I see. So, it is my it is our own defined parameter with which we want to describe the system and that means we should be able to choose parameters that describe the whole system properly.

So, whatever system we are concerned with. So, let us say if I am thinking about a queue. So, the queue at any frozen time is described by our system parameter which is the state. What is my state? Let us say how many customers are there inside the system that might be one state parameter. And then in the state, if I go means slightly into better nitty-gritty then it has to be described that at that point of time how many customers are there.

And if a customer is getting service then how much service he has already taken? So, this describes the whole state currently where the state is the means by which the system is observed. So, like that, as time changes you can easily see that the state will be changing either the number of customers will be changing or the amount of service the particular guy who is in server he is he has got that will be changing.

So, basically, it is the capturing in time of the evolution of the state that is exactly what is happening. So, therefore, these systems are exactly described by the associated timestamp and on that timestamp what is the observed state condition?

So, state versus time are the two things that capture the whole system. Now, what might happen? According to my description of my state and time, these two things one is probably the dependent variable one is the other one is the independent variable.

So, this time that might be also discrete or continuous. So, time can be discrete or continuous, similarly, the state also can be discrete or continuous. So, therefore, I will probably have four types of systems which has to be characterized differently.

So, it might be discrete time, it might be continuous time, it might be discrete state, it might be continuous state. Let us give an example of a discrete state number of customers if we say that is a countable thing. So, that is a discrete state if my state only consisted of the number of customers in the system then it is a discrete state.

If my state description also has the residual time that is there for the customer who is being served then it is a continuous one because the time residual time can take any real value therefore, of course, it has to be positive.

So, if it is greater than 0 any real value it can be ok. So, therefore, if I describe my state with that particular parameter that the residual time of the customer in the service, or if there are multiple servers then each customer what is the residual time? So, there will be a multi-dimensional continuous variable state or continuous state space I will be getting.

So, this is what will be happening, if my state description includes those kinds of things then I have a continuous state ok. Continuous states are difficult to handle because they can take infinite possibilities whereas, discrete states are a little bit easier to handle because they will only take it might be still infinite, but it is countably infinite it is all countable 1, 2, 3, 4 up to infinity probably ok.

So, that is the condition. So, state most of our analysis we will see that we will be concentrating on this discrete state and that is why probably we are emphasizing Poisson and this memory-less process In our last class we already demonstrated that if we take a Poisson thing, or if we take all exponential thing then the customer who is getting

residual service I can take him as a full customer this is something we have already discussed.

So, therefore, we really do not have to think about this partial customer or partial packet we do not have to. So, if I do that. So, mostly Markov has done that. So, any Markov chain generally it will be accounted for discrete time discrete state case the state space will become discrete ok. So, a number of customers let us say.

But the time can be anything now that depends on if in the timeline the events are occurring at particular chosen points only. So, suppose my time is slotted many of you might be discussing later on our means access protocol there is something called slotted aloha ok.

Where time actually is slotted only in the slot boundary you can do something. So, if time is slotted and only in slot boundary you can do something then that is called a discrete-time Markov process or discrete time associated, a chain we will talk about that chain and what we mean by a chain.

So, this is because of the representation graphical representation. So, associated things will be talked about as Discrete Time Markov Chain or discrete time Markov process. If the time is continuous like in our reality that happens then it is a Continuous Time Markov Chain or continuous-time Markov process.

So, what we will try to do now we will actually start describing our analysis requires this one, but without doing the description of this cannot be described. So, therefore, we will start with this one and you will later on see that we will also have application of this one.

So, our protocols will be analyzed by these things. So, if we talk about Ethernet protocol CSMA CD or if we talk about Wi-Fi protocol CSM CA that can be analyzed by discrete time we will also talk about that later on in the course.

But right now for trunk switch analysis, we need this continuous-time Markov chain or continuous-time Markov process. So, we will first give a brief description of the discrete-time Markov chain, and then from there, we will see how to proceed toward the continuous-time Markov chain. So, in our next class probably that will be our description or discussion.

Thank you.