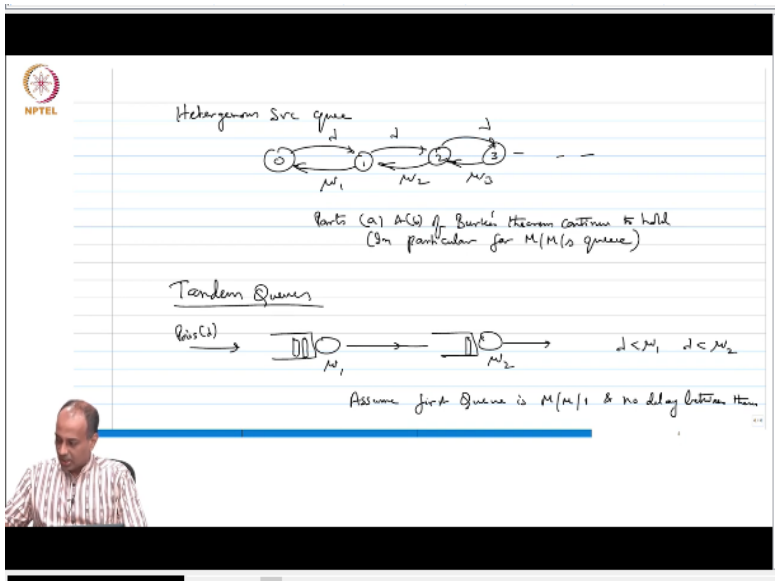
Stochastic Modeling and the Theory of Queues
Prof. Krishna Jagannathan
Department of Electrical Engineering
Indian Institute of Technology-Madras

Lecture-77
Burke's Theorem and the Tandem Queues-Part 2

(Refer Slide Time: 00:17)



Now I want to briefly discuss the issue of tandem queues. Tandem queues means you are putting one queue after another. So, here is a queue, so which has poisson arrivals at rate lambda and they are being served at exponential mu and the output. So, whenever these customers get out of the first queue they get into another queue which is another exponential server of rate mu 2 some other rate.

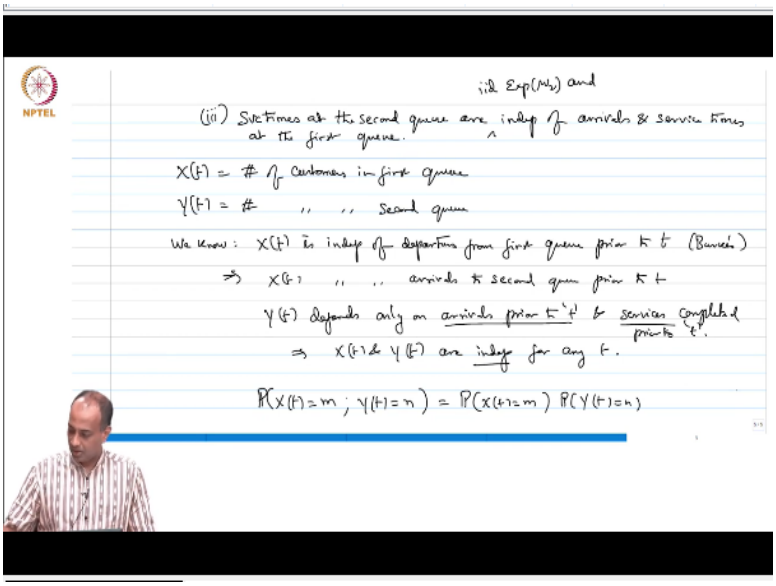
So, we are going to assume that lambda less than mu 1 and lambda less than mu 2. Mu 1 and mu 2 one of them can be bigger or smaller I do not care, but both mu 1 and mu 2 should be greater than lambda. This is well motivated in practice because often when you go to renew your passport or whatever you have to first go to one queue where they take your documents and you then get out of the queue get into another queue to actually apply for the passport or renewal of the passport or to renew your driver's license or whatever it is this kind of tandem queues are quite common as we know from practical experience.

So, here we assume that any customer who gets out of the first queue instantaneously goes to the second queue. There is no time delay in the intermediate, the moment a customer or a

packet gets out of the first queue he or she finds himself in the same for herself in the second queue. And both queues are exponential servers of rates mu 1 and mu 2 respectively. We are going to assume that see the first is an M/M/1 queue, this is an M/M/1 queue.

So, the first queue is an exponential server and the service times are independent of arrival times and the second queue there is no delay and no delay between queues, between them. So, a customer who leaves the first queue instantaneously joins the second queue.

(Refer Slide Time: 02:50)



Now we are also going to assume that so this is number 1, this is the second assumption, no delay between the queues, third assumption is service times at the second queue are independent of arrivals and service times at the first queue. So, the second queue is an iid exponential server, maybe I should say that also iid exponential mu 2 and are independent of arrivals and service time in the first queue.

o, if I am a customer I had some service time in the first queue and I get to the second queue my service time in the second queue is statistically independent of my service time in the first queue, all my arrival time in the first. So, these are assumptions that I am going to make. So, what happens in this case this system turns out to be fairly easy to analyze and that is because of Burke's theorem.

Now notice that the departure process of the first queue the first queue being an M/M/1 queue that departure process is also poisson of rate lambda, horizon of rate lambda. So, we are getting poisson inputs to the second queue and the second queue also has exponential service

rates. So, you may directly conclude that it is an M/M/1 queue but we are not there yet because you have to argue that the service times in the second queue are independent of the arrival times in this process in the arrival process to it.

Now we know that any given time t, so we have to argue that there is no correlation between the service times in the second queue and the arrival process to the second queue. Now we know that at any given time t the departure process is independent of the state of the system of the first queue and therefore you can argue that the arrival process here at any given time t is independent of anything that happened in this system. Also by assumption the service time here is independent of the service times here.

So, we can argue that the arrival process is independent of the service time in the second queue. So, you can argue that the second queue is a M/M/1 queue in a legitimate way. So, we are going to say now that so let us say X t is the number of customers in first queue and Y t is the number of customers in second queue at time t. So, we know that X t at time t is independent of departures from first queue prior to t.

So, X t is independent of departures from first queue prior to t, this is because of Burke's. Therefore X t is independent of arrivals to the second queue before time t. X t is independent of arrivals because the departures from queue 1 prior to time t are in fact the arrivals to say queue 2 the second queue prior to time t. Depend of arrivals to second queue prior to t, because there is no we are not wasting any time waiting in between.

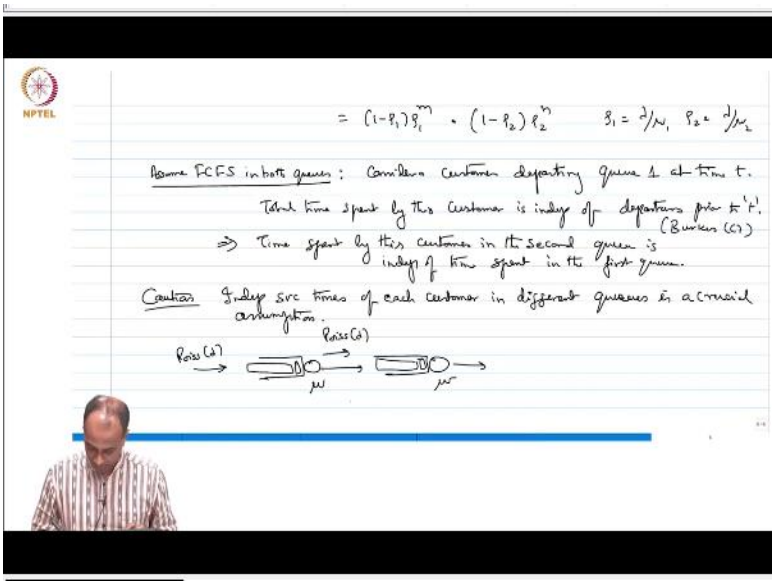
Now Y t, this is true for any t. Now Y t depends only on arrivals prior to t and services completed prior to t. So, Y t, the number of customers in the second queue at any time t will depend on the arrivals that happened before t and all the services that happened before t. We have already argued that the arrivals prior to t which are the departures prior to t from X t, so Y of t is dependent only on arrivals prior to t which are independent of X t.

And services prior to t the service times prior to t which are also independent of X t. Because X t the service times in the second queue are independent of the service times and arrival times in the first queue. So, both arrivals prior to t and the services completed prior to t are independent of X t and Y t depends on only these 2 things. So, you can argue that X t and Y t are independent random variables for any t.

So, we have argued 2 things; first that the second queue is also an M/M/1 queue because you have passions arrivals and exponential services. And the second queue the service times are independent of the arrival times because the arrival times only depend on the service times and arrival times the first queue and which are independent of service times in the second queue.

So, we have argued that the second queue is also M/M/1 and we have argued that X t and Y t the state of the 2 systems are in fact independent. So, we can easily argue that I mean let us also assume that we have maybe we can assume FCFS if you want things to be even simpler. So, you have seen what is now X t and Y t are independent random variables. So, you can argue that probably that X t = m and Y t = n, you can write as P x t = m times P y t = n. This is because of independence at any given time.

(Refer Slide Time: 11:55)



And probability X t = m is nothing but 1 - rho 1, rho 1 to the m and probability Y t = n Y 1 - rho 2 rho 2 to the n where mu 1 and rho 2 is equal to lambda over mu 2, both are assumed to be less than 1. So, it is easy to find the joint distribution of X t and Y t at any given time, it is just a product of the usual M/M/1 distributions. Next list if you assume so this so far everything we have said is valid in full generality, we have not assumed any FCFS.

Now if you assume FCFS in both queues now you consider a customer who departing queue 1 at time t. Now the arrival time of this customer and therefore the total system the time spent by this customer in queue 1 is independent of departures prior to t. So, the total time spent by

this customer is independent of departures prior to t this is because of again Burke's theorem part c this is.

Now we are using FCFS and part c of Burke's theorem. So, we are saying that total time spent by the customer who departs at time t departs queue 1 at time t is independent of departures prior to t and this departures prior to t are in fact the arrivals into the second queuing system. So, the time spent by this customer in the second queuing system will be independent of the time spent by this customer in the first queuing system.

This implies time spent by the by this customer in the second queue is independent of time spent by the same customer in the first queue. So, if you want to look at the total time spent by the customer in the first queue plus the second queue it will be a sum of 2 independent random variables only for FCFS. So, if you have FCFS the customer who departs the first queue at time t the system time in queue 1 alone is independent of departures prior to t.

But the departures prior to the customer joins the second queue will determine his waiting time in the second queue, but we know from Burke's theorem part c that the departures prior to t are independent of the customers total time in the first queue. So, the time spent by the customer in the first queue is independent of the time that the customer will spend in the second queue. So, the total time spent by the customer in queue 1 plus queue 2 will be a sum of 2 independent random variables.

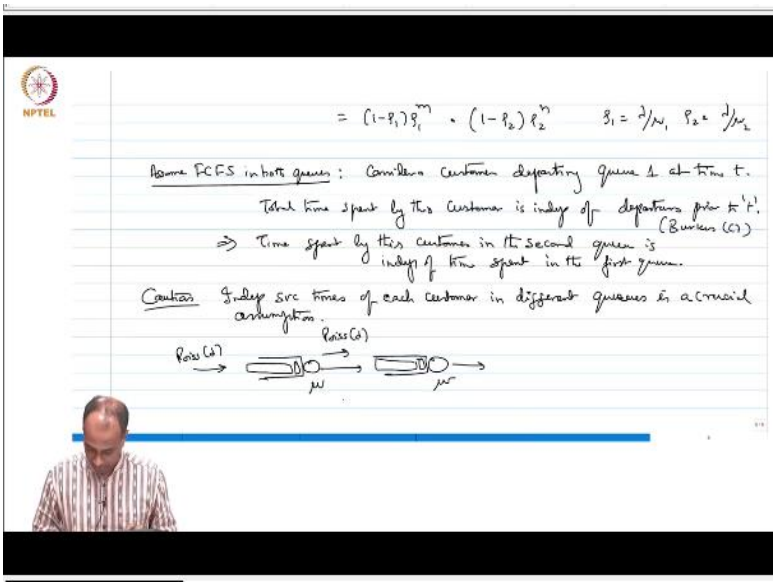
That is what this part is saying using Burke's theorem part c. We already know that the time spent by a customer in an M/M/1 queue is an exponentially distributed random variable with parameter mu – lambda. So, what happens is the when the customer enters the first queue she spends exponential amount of time with parameter mu 1 - lambda and then goes to the second queue and spends an independent exponential time with parameter mu 2 - lambda and then leaves the system.

And then if you want you can add one more queue, you can add a third tandem queue and all these arguments will repeat. So, if you have a whole bunch of tandem queues with exponential service times you can make 2 important conclusions. The one is that maybe 3 important conclusions, all of them behave like independent M/M/1 queues.

In the sense that at any given time t the joint distribution will be like a product of the corresponding M/M/1 queue's distributions of course the queuing processes are not independent, because you can argue that if X t if there are never any customers in the first queues of course there would not be any customers in the second queue, it is not as so these tandem queues are like independent M/M/1 queues that is not what I am saying.

I am saying that at any given time t the distribution products out. So, we can say this even for 3 tandem queues or 4 tandem queues or whatever you want and in the FCFS case the total time spent by a customer in the first queue and then the second queue are independent and you can add them up as independent random variables to get the total time spent in the system and that is because of Burke's theorem part c. So, this is quite powerful. See there is one caution I want to give you that it is crucial.

(Refer Slide Time: 18:10)



So, this independent service times in 2 queues is very crucial, service times of each customer in different queues is very crucial. So, what we are saying is that if I spend a certain amount of time in queue 1 and certain amount of time in queue 2 I am assuming that these 2 random variables are independent. If this is not the case then all of what we are saying will actually break down.

Let me show you one example where the service times are not independent I am going to show you a slightly extreme kind of an example where let us say they have 2 queues I have poisson arrivals of rate lambda; this is exponential service of rate mu. I have another queue also of rate mu mu 1 = mu 2 let us say. Now what I am going to say is that so in this case

well so far I have assumed that if I am a packet of some size here and I get served here for some time I take an independent avatar so to speak in the second queue.

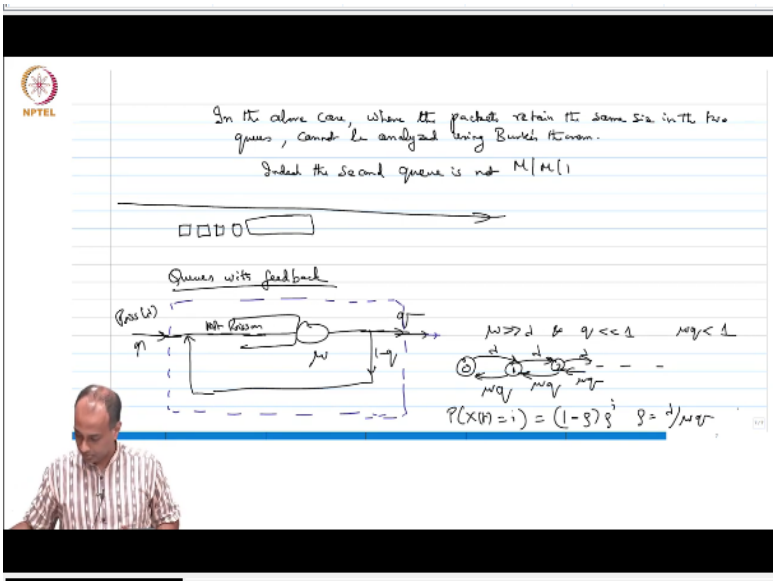
So, what I mean by independent service time is that a packet or a customer in the first queue takes a certain amount of time and a second queue takes a certain different amount of time which is independent of the first. So, I do not retain an identity of the amount of service. So, for example this could be justified if you are verifying documents in one counter and going to the next counter to renew your passport or whatever this assumption could maybe it is justified**.**

But in a communication network this is not at all justified because I do not take different avatars I mean it is the same packet after this customer me is not a person going to a passport office let us say I am a communication packet I am like a packet of a certain size then I am not going to take different avatars. So, a packet which is small here will continue to be small here, let us say a certain amount of bits and a packet which is fat will continue to be fat.

So, this is an extreme case where the identity the amount of service time from one queue to another it is not only not independent I am actually maintaining it to be the same. In this kind of a scenario you cannot argue that Burke's theorem holds, Burke's theorem does not hold at all and in fact the second queue will not be M/M/1. First queue is M/M/1 of course. So, where do things break down?

So, the output process of this is still poison, because the output of an M/M/1 queue is always poison and the second queue is of course an exponential server, but I am arguing that in the case when these packets retain their sizes or identities or their service times the second queue is not an M/M/1 queue and you cannot use Burke's theorem. Now why is that the case that is because there will be heavy correlation between the arrival process here and the size of the packet that is coming in. In this case let me write this down.

(Refer Slide Time: 21:59)



In the above case where the packets I am now calling the customers as packets because I am saying they have a particular size. So, the service times in the 2 nodes; here the 2 routers or whatever they are not independent; they are in fact the same. Well the packets retain the same size in the 2 queues cannot be analyzed using Burke's theorem. Why is that indeed the second queue is not M/M/1, and you know why?

See the arrival process is poisson, the arrival process to the second queue is poisson and the service times are exponential, but the arrival times is not independent of the size of the packet that is coming, see this is an important assumption in the M/M/1 queue. In an M/M/1 queue what do we assume the arrival processes are poisson and the service time of each customer is independent of the arrival process is what we assume in an M/M/1 queue and that is not true here?

You know why because see just think of what happens, let us say I have some big packet, this is a big packet, let us say this is my time and this is a big packet. He gets served in the first queue, so this guy over here is getting served whenever he is getting slowly served, so when slowly this job is being eaten away by the first server there will be a lot of other smaller jobs that would keep coming at rate lambda.

So, when the first job is getting served there will be a lot of smaller jobs that would come to the first queue, let me typically smaller job. So, when you have an atypically large job that is sitting at the first queue there will be a lot of small jobs that come. So, whenever the big file

finishes service at the first queue and gets to the second queue it will be immediately followed by a number of small packets.

So, what does that mean? When I see a big packet I can imagine that there will be number of arrivals after that in succession in that for the second queue. So, what we are saying is that there is no I mean if you look at the size of the arrival process to the second queue still poisson, but if I look at the file if I see a big file I know that there will be a lot more smaller files that come after that.

That is because they have been waiting behind this big file. So, this is somewhat like I did not make this very precise, it is somewhat like if you have a big truck moving on a small road there will be a number of cars behind it. This is like a slow truck effect, because all these cars have just come and started waiting behind the truck because the truck is moving very slowly something like that happens out here, you can simulate this and see if you wish and it was not very, very mathematically precise but this is exactly what happens.

So, there is a correlation between the sizes of the files and their arrival times that is what happens in the second queue. So, although the arrival process is poisson it is not an M/M/1 queue that is because it is a correlation between the sizes of the files and the arrival process. So, in fact this kind of a tandem queue cannot be analyzed under works and may not be easily analyzed at because there is such a strong correlation between the file sizes in the arrival process. So, this is not an easy process to analyze.

So, this is a word of caution that this Burke's theorem should be used very carefully, it crucially depends on the service times in the 2 queues being independent random variables. Another word of caution is that if you have cycles this sort of breaks down. Let us say so if you have tandem queues where one queues output is feeding the other the service time in the second output is independent of the first and the second output is feeding the third one where again there is independence across service times then you can use Burke's theorem tandem queues analysis; then you will have independence of the state and all that.

But the moment you start feeding back, so feed forward is perfectly. If you have feedback, so if the output of one queue feeds back into its own input or another queues input then you will have trouble, let me show you. Let us say feedback, let us say I have this queue, so I have an

arrival process of poisson rate lambda; they enter the system and they get served and then I split what do I do?

I split a fraction of 1 - q and q with probability q a served customer departs the system and with probability 1 - q a served customer instantaneously joins the q. So, I complete service, I toss a coin independently across other, so this q can be thought of as a coin toss. So, moment I finish service I toss a coin if it turns out like with probability q I leave with probably 1 - q I instantaneously joined back the q.

And this q this coin toss process is independent of everything else, independent of the arrival process, service process and it is independent across customers and so on. So, in this system it is a little bit problematic, this is not really an M/M/1 queue. Why if you look at this let us say if mu is much, much bigger than lambda and let us say q is much, much smaller than 1. Then what happens is that let us say the system is empty, a customer would come lambda is much smaller than mu, so mu is very fast.

So, it would get served but it is very high probability that the customer will come back and again get served and come back, again get served and come back and after maybe let us say if q is 0.01 then I would come back about 99 times and 100 time I would leave. So, the timeline would come like a customer comes very well 99 times he comes back and goes. And after a very long time a new exogenous arrival will come, again the same thing sort of thing will happen.

So, if you look at the process here is poisson that is what we know is poisson. But the process here is not poison, it would be very bursty. So, the process that is actually entering the queue is not at all poisson because it will have a number of repeated entries and then nobody the number of repeated entries and nobody and so on. So, it is not a poison input to the queue at all in this scenario, but if you can there is a way to analyze this kind of a system; let us say you put it inside a box like so.

Now if you look at the Markov chain corresponding to the number of customers inside this box this 1 - q the customer joining back it is like a self transition and will not be seen as a change in the state of the system at all. So, the Markov chain for the number of customers

inside the box will still look like it is a Birth-Death chain except the rate at which this will be like mu q.

Assuming that mu q is less than 1 and this can be made into a positive recurrent chain, let us assume that mu q is less than 1; mu is very large compared to lambda and q is very small but mu q is less than 1 and so on dot, dot, dot. So, the state of the system inside this box still behaves like it is an M/M/1 queue with parameter lambda over mu q as the rho factor. So, if you look at X t as the number of customers in the system it will still satisfy 1 - rho times rho to the i where rho is equal to lambda over mu q.

Although the queue itself forget the box now, if you look at the queue itself it's not at all an M/M/1 queue it has burst arrivals because the same guy keeps coming back, but nevertheless it is a Markov process that we can very well understand as a Birth-Death chain and you can solve for its occupancy probability. So, that is just I mean it is an interesting system to consider.

So, anyway so you can think of this now that we have studied tandem queues where you feed forward and you can consider them as not as independent M/M/1 queues but any given time they behave like they are independent M/M/1 queues and then you have these queues with feedback. So, you could potentially even take with probability P i take the output of q 1 and send it to q 3 with probability some other probability I send you to q 4 and bla, bla, bla.

I can make a big network of queues like this there are some exogenous inputs, some departures which leave the system for good but some departures get routed from one queue to another queue with some probability and so on. These kind of systems can be studied they are known as Jackson networks and that will be the topic of our discussion in the next module. So, I will stop here, thank you.