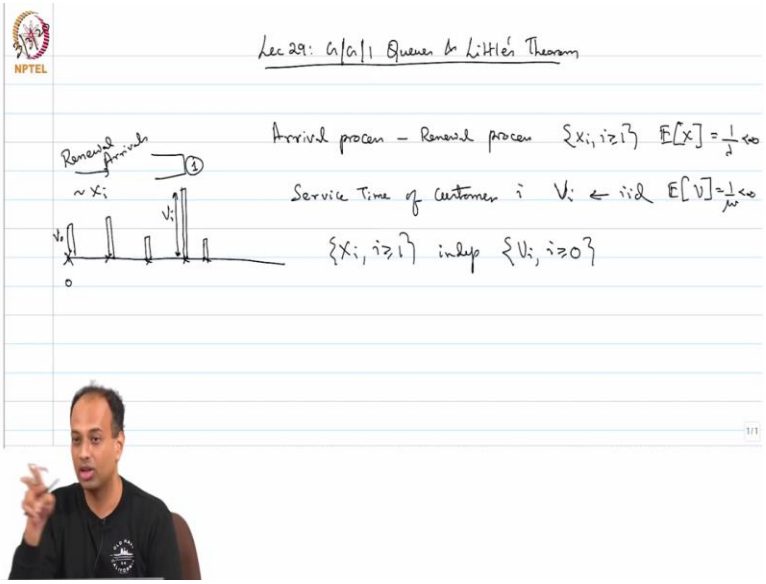
Stochastic Modeling and the Theory of Queues
Prof. Krishna Jagannathan
Department of Electrical Engineering
Indian Institute of Technology – Madras

Lecture –35
G/G/1 Queue and Little's Theorem

(Refer Slide Time: 00:13)



Today we will discuss the G / G / 1 Queue and a very important result in queuing known as Little's theorem which relates the average occupancy of a queuing system to the average waiting time in the queuing system. So, we will discuss the G / G / 1 Queue a little bit and we will prove Little's theorem using the renewal reward approach that we have developed for the G / G / 1 Queue.

Although, the Little's theorem itself is not just limited to G / G / 1 Queue so it is actually applicable even to multi server systems as we will see, but we will prove it only for G / G / 1 that this today's agenda. So, you guys already know what the G / G / 1 Queue is. So, this is a server and this is a queue. You have renewal arrivals according to some distribution x. So, x i is of inter arrivals.

And the service times of each job are also iid. So, that is actually what is usually known as GI / GI / 1 Queue. You have independence between consecutive service time independence between consecutive arrival times. So, some people write GI / GI / 1, but usually G / G / 1

means without any further qualifications usually means GI / GI / 1. So, you can look at iid so arrival times arrival process is a renewal process with inter arrival time x i's.

And expected x let us say 1 / lambda, lambda is the arrival rate and this is assumed to be finite and the service process so each customer whenever this renewal process ticks there is a customer that arrives. So, if you look at the time axis let us say this is time 0. So, you have some renewal arrivals and whenever this arrival occurs there is a customer who comes to the system.

And each customer brings with him or her an inherent service time. So, we assume that the server works at a constant rate let us say 1 and you can think of this customers as bringing with them a certain file size or job size or whatever. So, some could be large and some could be small and what we are saying is that these random variables are also iid this file size or customer service time random variables are also iid.

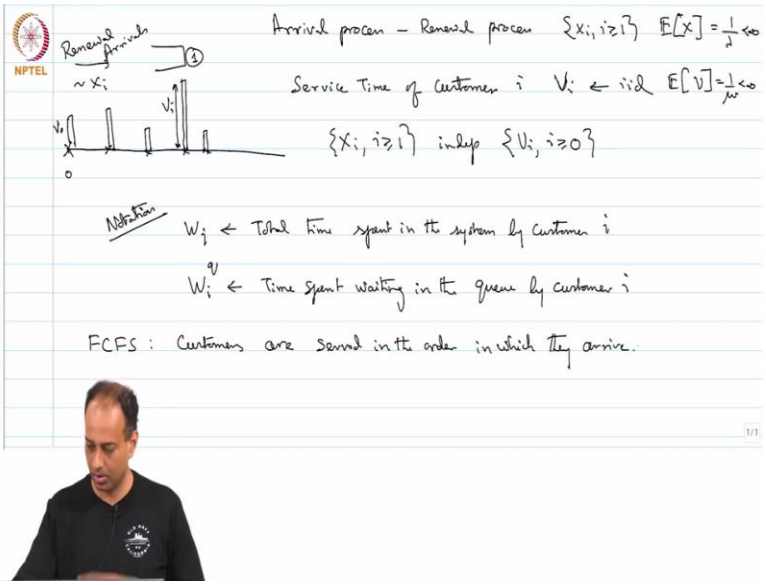
So, these are the file sizes or let us say v i so you can say that the servers serves at a constant rate 1 and each file of size V i will take an amount of time V i to get served at the server. So, we assume that the service time of customer i is also v i which are also iid and we will say x i so this sequence x i is independent of this sequence V i. Well this sequence V i I will write i greater than or equal to 0 because for the G / G / 1 Queue convenient convention is to start ticking time when the first arrival comes to the system.

The system is empty for a very long time there is nobody at all and then suddenly the first arrival comes and you start counting time at that point. So, what you will do is you can take the first arrival at 0. So, in some sense it is renewal process after time 0 there is a first arrival at time 0 always because you are starting your clock at that time. So, this you can think of as customer with file size V 0 who comes to the system at time 0.

Starts getting served and then the subsequent arrivals are renewal arrivals and each of course brings in a job size or file size or customer service time V i. So, this x i are iid, V i are iid again we can write expectation of V as 1 over mu it is also assumed to be finite though and x i are independent of V i. So, this is the model of a G / G / 1 Queue. So, the arrival distribution can be anything, inter arrival distribution can be anything, service time distribution can be anything.

They just cannot be correlated among themselves or across the x i and V i that is the G / G / 1 Queue for you.

(Refer Slide Time: 06:20)

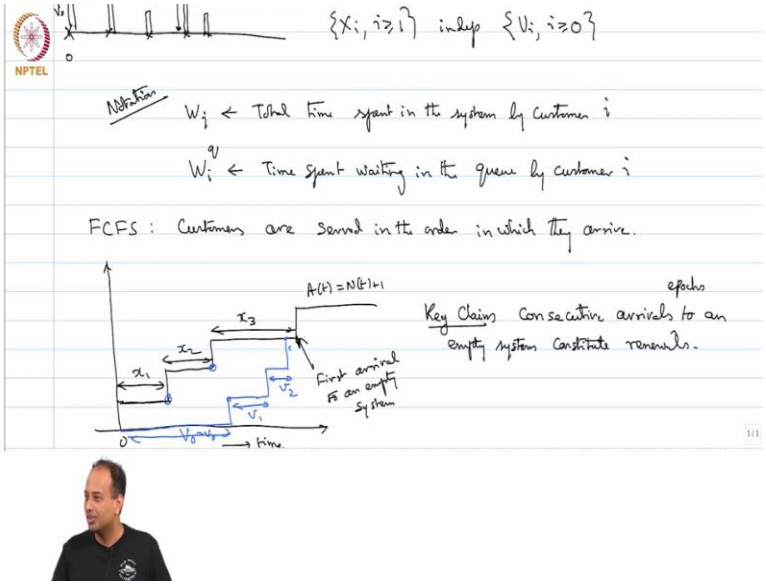


Now, we are interested in looking at the waiting times of these files or these customers in queue. The usual notation is W i is taken as the notation for total notation, total time spent in the system by customer i and the book uses W i q. So, these are random variables so these big capital letters. Just in case you are confused W i q is the time spent waiting in the queue by customer i.

So, what is the difference between W i and W i q? W i is the total from the time customer i arrives at the system, wait, get into service, finish service and get out that is the total system time. W i q is simply the time spent waiting not including the time getting served. So, W i is basically the W i q plus your own service time you spent getting served yourself. Now, we are going to do I have to specify the service discipline.

I am going to do only first come first served in class, but I will also point out why this FCFS is not I mean much of what we are going to say in today's lecture we will hold for other service disciplines as well, but FCFS is probably the most intuitive and most common. So, this is first come first served. So, customers are served in the order in which they arrive. So, person who came earlier will be served earlier.

(Refer Slide Time: 09:04)



This is not of course necessary you can do something else also. I am just going to analyze this situation FCFS situation. So, please remember that there is the time begins when the 0th customer comes at time 0 with some file size V 0 that is when we are starting time. So, I want to just plot for you the evolution of this FCFS queue G / G / 1 queue. So, the first arrival happens at time 0.

So, let me plot the arrival process has 1 + N t where N t is the renewal process. So, it starts at time 0 and there is an arrival here let us say. So, this is just the arrival process I am plotting the arrival process. So, let us say this one more arrival and so on. So, this I can call it as arrival process this is time of course I am going to call it A of t = N of t + 1 because N of t is the arrival process renewal process N of t.

Why am I adding + 1? There is arrival at 0 always after that you are counting renewals. Now, I also want to plot the departure process. So, this is the arrival process then this is FCFS and I want to plot the departure process. So, the customer who arrives at 0 starts getting served at 0 he comes to an empty system. Now how long will he be in the system the first customer to come at 0 his service time which is V 0.

So, his departure will occur at time V 0 so this let us say that is my departure process I am drawing this in blue. So, this length if you look at it right this length will be equal to this length will be V 0. Of course, these lengths are your x i may be here I should not write apologies so these are realizations. So, maybe you should write so this is like x 1 is equal to little x 1 this is x 2 that is x 3 and so on these are the inter arrival times.

And this V 0 there is some little V 0 and when this guy leaves the second person gets into service. So, the second customer who gets into service at this point is the person who arrived over here is that right. So, that person get serve for some time and then he leaves and then the person who came here gets into service. The second arrival not counting the arrival at 0 gets into service and in this picture let us say he finish a service.

So, in my picture so this guy is v 1, this guy is v 2 this interval is v 2. So, v 1 corresponds to the service time of the customer who entered here and v 2 corresponds to the service time of the customer who entered here. So, at this point how many people are there in the system? System has emptied and then in this small duration system is empty and then there is another arrival and so on (()) (14:03) sort of thing repeats.

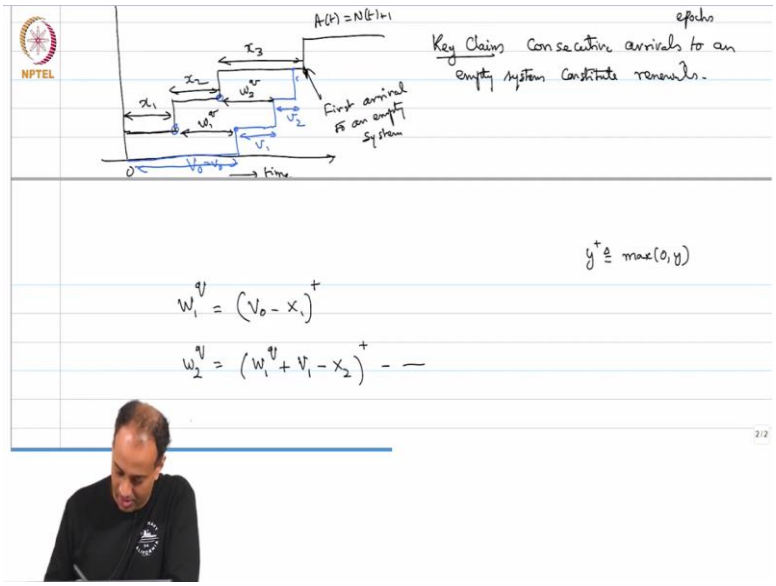
Now a key fact that we will argue prove at least prove using pictures we can make this very rigorous easily is that consecutive arrivals epochs to an empty system constitute renewals. This is a key fact. So, what do I mean? So, if you look at this picture at time 0 you had an arrival to a empty system and what is the next arrivals that occurs to an empty system this guy.

This is the first arrival to an empty system of course not counting the 0th arrival to the empty system. The time at 0 is the beginning of time a bunch of arrivals come and they leave and then there is a another arrival to an empty system. Now of course the similar sort of thing plays out. This guys getting served more people will come and so on and let us say this system empties again and one more time there will be an arrival to the empty system.

What we are saying is that these durations constitutes renewals so that is what we are going to show. This is not all the difficult to see because in some sense if your system emptied out statistically what is going to happen after if you have an arrival to an empty system statistically what is going to play out again is the same as what played out starting a $t = 0$ you again have some arrivals which have inter arrival time the renewals and service times are iid.

I am not saying they will be equal of course they will be different, but statistically they will be the same and statistically they will also be independent because all these x_i and v_i are independent you see what I mean so that is what we have to argue.

(Refer Slide Time: 17:08)



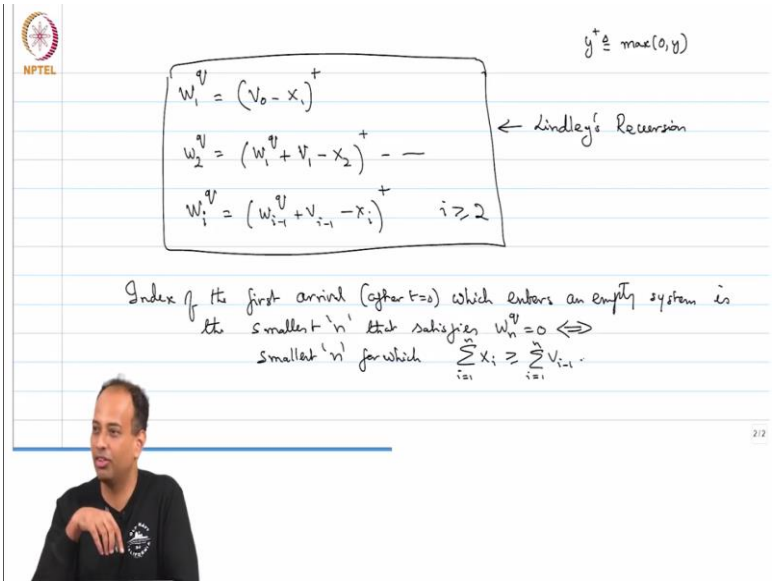
Now to understand why this is true I want to be able to track how long these customers wait before they get into service. Of course, the arrival at 0 waits for 0 time. The second arrival which occurs after t = 0 how long does that person wait? The x 1 arrival so to speak waits till the time 0 arrival complete service. So, this is the amount of time that the customer number 1 waits in queue before getting into service.

So, this should be w 1 q. So, in this picture I can write w 1 q = I can write this in terms of random variables. I have drawn specific realizations I can write w 1 q = how much v 0 – x 1, but see in my picture v 0 is bigger than x 1, but if the amount of time if x 1 the first arrival took place after v 0 completed then what happens the waiting time will again be 0. So, if v 0 – x 1 is negative then I should count this as 0 I should not count negative time.

So, I can put plus max of whatever this is, 0. So, basically I am using the notation y + = max of 0, y. Now what happens to the second customer? The x 2 arrival so to speak waits for that much time. So, what is w 2 q can you tell me some of those equations? I would like it in terms of the previous waiting time so it will be w 1 q + v 1 – x 2. So, what am I saying so it is w 2 q is simply it is the time that it arrived.

So, this total length is w 1 q + v 1 – x 2 will be w 2 q. So, let me just write it out. I have written it is w 1 q + v 1 – x 2. The plus will come for the same reason if this is negative you will have to count it as 0. So, you can generally say you can keep doing this.

(Refer Slide Time: 20:24)



You can write for the Ith customer you can write w i q = w i – 1 q + v i – 1 – x i whole plus. If this turned out to be negative then you will put 0. This is true for i greater than or equal to 2. So, if you take so this bunch of equations actually I can actually just delete the second equation it is just a consequence of it is contained in the third. So, this bunch of recursive equations are known as Lindley's Recursion.

It is a recursive equation you can write to characterize the waiting time of the Ith customer in terms of the waiting time of the i – 1 customer and the x i and v i. See this x i and v i are given to you. I mean those distributions are known to you and you will be able to write these waiting time in terms of this x i and v i and the previous waiting time. This is a first come first serve G / G / 1 Queue.

Now, we want to see the 0 that arrival at time 0 had 0 waiting time then wait at all. Now, what is the index of the next arrival for which faces 0 waiting time. So, in our picture for example the x 1 arrival the first arrival not counting the arrival at 0 has some positive waiting time w 1 q and then w 2 q was also positive and the third arrival which happened to an empty system had 0 waiting time.

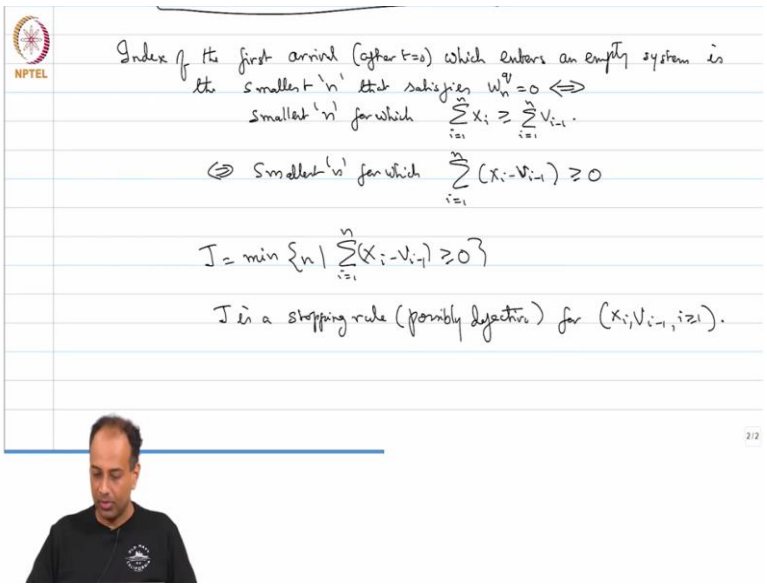
If you think about why this happened? What is it about that the third arrival in terms of the x i and v i that helped it to face 0 waiting time. See, it is the first arrival for which the sum of the x i is greater than or equal to the sum of the previous v i. See for the first arrival v 0 was bigger than x 1. For the second arrival also w 2 q was positive because v 0 + v 1 was bigger than x 1 + x 2.

But for the third arrival that x 3 is quite large in the way I have drawn it. It turns out that the x 1 + x 2 + x 3 which is the epoch x 3 of the third arrival epoch is larger than the sum of the service time of all the customers who have come before. In fact you can see this from the Lindley's Recursion w i q = 0 for the first time what is the smallest i for which w 1 q = 0 will also be the first index for which the sum of x i is still that particular index is exceeds the sum of the v i – 1 I guess.

So, I am going to write it like this so index of the first arrival so after t = 0 which enters empty system is the smallest n that satisfies w n q = 0. This is the thing about the first arrival sees an empty system this is equivalent to the smallest n for which sum over i = 1 to n x i is greater than or equal to sum over i = 1 to n v i – 1. In the picture I have drawn this n is 3. Now, I will argue so what does this depend on?

So you think about this. So, you have a bunch of random variables further over the queuing system, you have a bunch of iid random variable x i. We have given a bunch of iid random variables v i – 1 and I am asking you keeping looking at these x i and v i and tell me the first time that sum over x i exceeds sum over v i – 1. Now what does this sound like? This sounds like something we have already studied.

(Refer Slide Time: 26:25)



So, let me put it this way so I can write this as the smallest n equivalently the smallest n for which sum over i = 1 to n x i – v i – 1 is greater than or equal to 0. See you call x i – v i – 1 as some other variable y i they are also iid. So, we are looking at the first time that these y i

could be positive or negative I am looking at the first n such that sum over y i is now negative. What does this sound like?

See this is like a stopping rule except that there are two sequences of random variables so you know what sequence of I mean stopping rule J for a sequence of random variable x i you look at each n you look at just the first n realization and decide to stop or not. Now, I am going to look at not just x i I have a bunch of x i, I have a bunch of v i and I am going to look at both and I am going to stop whenever sum over x i exceeds greater than or equal to sum over v i – 1.

So, I am write a stopping rule J as the minimum of n such that. Now, clearly the event you can argue you are only looking at x 1 through whether you stop at n depends only on x 1 through x n and v 0 through v n – 1.So, is J is a stopping rule it could be defective. It is not clear that you will eventually see a arrival to an empty system what if queue keeps building up or something.

It could be defective possibly defective for not just x i, but x i v i – 1. So, you can define this is like a slightly generalized version of stopping rule where you do not have just one sequence you have two sequences. It does not matter you are only looking at so whether you stop a time in or not is dependent only on the information available at that time you are not looking ahead.

You are looking at x 1 through x n and v 0 through v n – 1. Remember, J is not a time J is an index that is why I am writing stopping rule instead of saying stopping time. It is the index of the arrival it is not actual time. Now, what we will see is that J is a genuine stopping rule not a defective stopping rule if the average inter arrival duration is greater than the average service duration if expected x is greater than expected v.

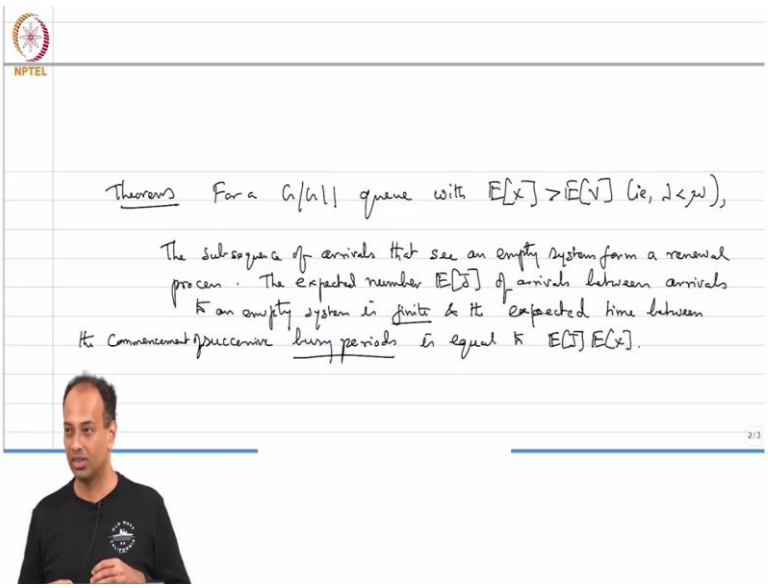
Then J is a legitimate stopping rule which means that if you will stop in finite you will always have an arrival with probability 1 which enters in empty system. Furthermore, if expected x is greater than expected v you can show that expected J is finite slightly non trivial, but you can show it and what we can also show is that once the stopping rule realizes as J = some n then these subsequent arrivals let us say condition of J = n.

The subsequent inter arrival time will be x n x x n + 1 etcetera in our case n = 3 then the subsequent inter arrival times are x n x n + 1 etcetera and this subsequent service times are v n – 1 v n etcetera and again Lindley's equation can be written for the subsequent waiting time and all that. Now, you can also argue that condition on J = n what are the inter arrival times x n x n + 1 etcetera which are independent of the previous inter arrival times in the previous busy period so to speak when the server were busy.

And likewise for the v n v n – 1 v n + 1 etcetera are independent of the previous service times and all these for every n your conditioning on J = n these random variables x n x n + 1 etcetera and v n etcetera are independent of J and independent of the previous x i and v i in the previous busy period. So, I have argued independence and identical distribution condition on J = n, but this is true for every n.

So, you are basically argued that these are renewals. What do I mean by these are renewals? The times at which that are consecutive customer entries to a empty system these constitute renewals.

(Refer Slide Time: 32:37)



So, we will state this as a theorem for G / G / 1 Queue with expected x greater than expected v i. e lambda less than mu. Thus, subsequence of that see an empty system in the renewal process. The expected number of arrivals between arrivals to an empty system is finite and the expected time between successive busy periods between the commencement of successive busy periods is equal to expected J times expected x.

This is the expected inter renewal duration. So, I have used the term here busy periods. So, you go back to this picture so this particular picture shows one busy period. So, you have an arrival to a empty system then there is a whole queue build up wait, wait, wait everybody leaves and then the system empties then there is another arrival. So, I am saying these successive intervals constitute renewal period and this is one busy period basically.

So, what we are saying is that so we are saying two things let us go back to the theorem now. We are saying that the subsequence that see an empty system form a renewal process. So, what do I mean by subsequence? You have the arrival sequence x 1, x 2 this is the renewal process to begin with, but not all of them see a empty system. There is a subsequence that sees an empty system.

So, in our picture the first arrival of course is an empty system I mean first arrival at 0th arrival sees an empty system then the third arrival sees an empty system then maybe the 7th and 13th or whatever see an empty system. We are saying that particular sequence that particular subsequence of this x i itself constitute a renewal process and we are also saying that J which is the stopping rule which is the minimum x which is the first index for which you see a empty system.

This J is a legitimate stopping rule with finite expectation that we have not proved. It is not difficult to prove it I think there is an exercise in your book it uses a truncation argument we have to use inter arrival distribution which are bounded prove it for that and then use a truncation argument for proving expected J is finite and then the expected real time, the continuous time between one arrives which is an empty system and the next arrival which is an empty system.

The expected time will simply it is like expected S J which is just by Wald. Expected J times expected x which is also be finite. If expectation of x is greater than expectation of b that is the arrival rate should be less than the service rate. Lambda is the arrival rate which is 1 over expected x and mu is 1 over expected v so lambda is less than v then we are guaranteed that you will have infinitely many of these entries to empty systems.

And the expected duration between these renewals these expected renewal time will also be finite which is just expected J time expected x. Now we will use this bigger renewal process

this what I call the subsequence what sees empty systems to prove Little’s theorem that will be a discussion next class.