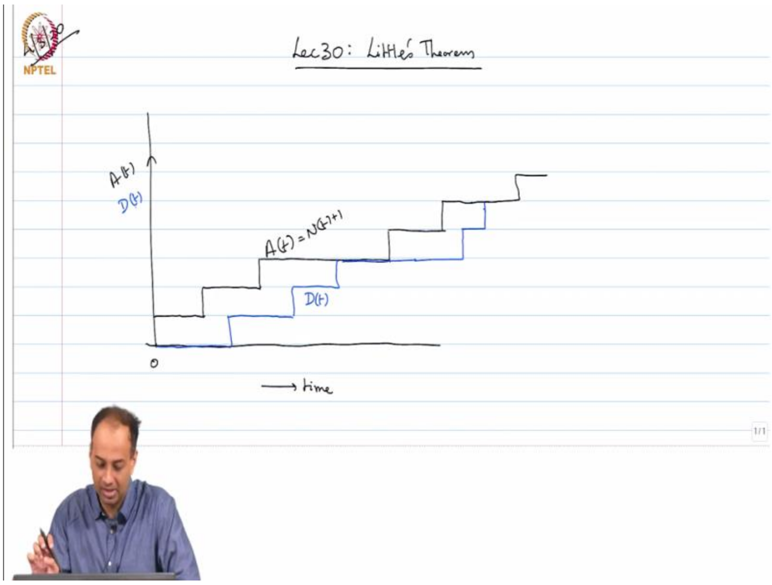
Stochastic Modeling and the Theory of Queues
Prof. Krishna Jagannathan
Department of Electrical Engineering
Indian Institute of Technology – Madras

Lecture –36
Little's Theorem

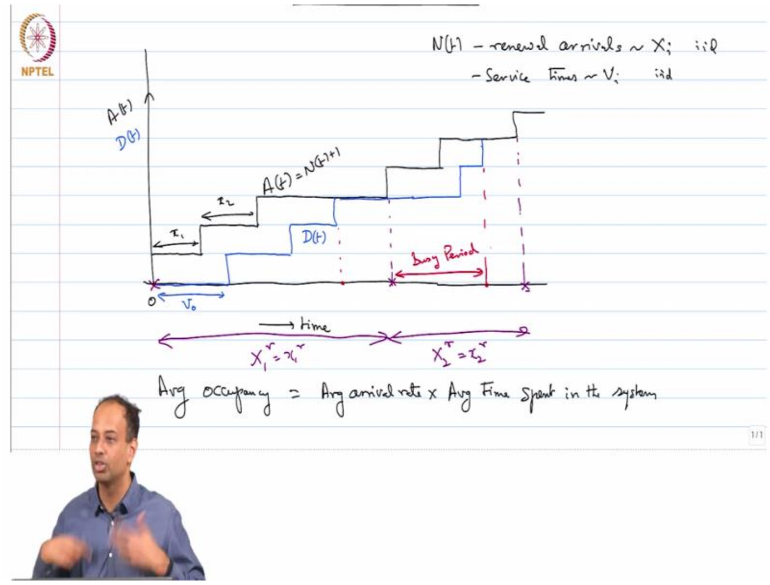
(Refer Slide Time: 00:14)



Lec30: Little's Theorem

Welcome back. Good morning. Today, we will discuss and prove a very important result in queuing theory known as Little’s theorem. So, Little’s theorem relates the average number of customers in a queuing systems to the average waiting time in the queuing system.

(Refer Slide Time: 00:32)



N(t) – renewal arrivals ~ X_i iid
– Service times ~ V_i iid

Avg occupancy = Avg arrival rate x Avg time spent in the system

Roughly it says that average occupancy is equal to average arrival rate times average time spent in the system. So, this is roughly this is the intuitive this is what it says. So, we have to make precise what we mean by average occupancy, average waiting time etcetera, average time maintenance system etcetera.

The way you prove this is using renewal reward approach in particular we have to use the renewal picture we put out yesterday. We said that in G / G/ 1 Queue. We looked at FCFS, but this is not just only for FCFS. We said that consecutive arrivals to an empty system constitute renewals. So, in this picture if you look at this picture that I have drawn I have drawn this black step function is the arrivals.

There is an arrival at 0 the system starts at 0 in this arrival to an empty system and then this black step is basically denoting the arrival process A of t which is N of t + 1. N of t is the renewal process of arrivals and these are distributed as some x i and service times are distributed as v i. These are iid and x i and v i are independent. Now I have drawn the black step which I said is the arrival process and the blue curve is the departure process.

So, at any given time t A of t denotes the total number of arrivals that have come to the system and D of t denotes the total number of departures from the system. So, this is your little x 1 the realization of x 1, this is x 2 and so on and this is your v 0. Now the renewal instances are can you mark in this. So, this is the 0 of course is arrival to a empty system and then this arrival if you look at that is an arrival to an empty system.

So, these purple crosses are arrivals to an empty system. So, in the picture the third arrival this is epoch of the third arrival and this one is the epoch of the fifth arrival is not it. So, this is the subsequence which sees empty system so we said yesterday that this subsequence of these arrivals with (()) (04:29) empty system constitute renewals. So, what we are saying is that these durations and similarly subsequent arrivals to empty systems are iid basically.

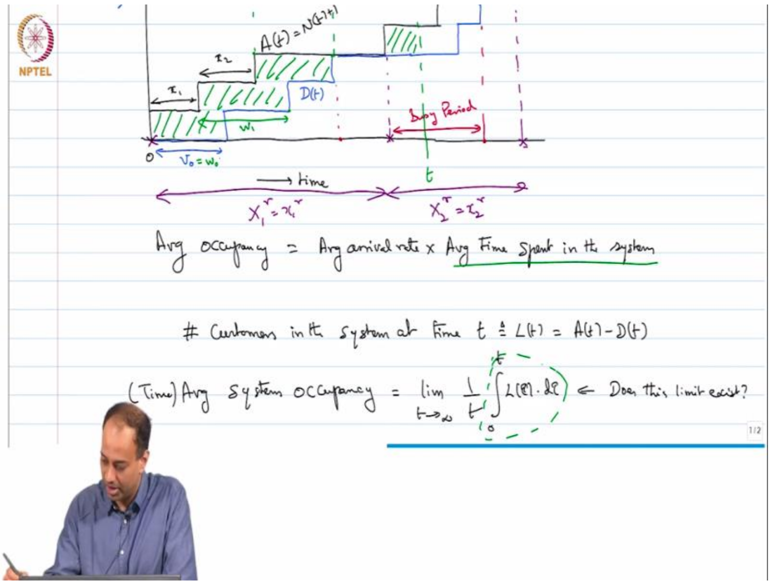
So, we can denote these by x 1 r x 2 r you know this is a different notation. So, this is the inter renewal periods of the subsequences. So, these are renewal instances I am putting r to distinguish them from the arrival process which is also renewals. Arrivals are x 1, x 2 etcetera this x 1 r, x 2 r etcetera are these durations between arrivals to empty system and each of these renewal intervals constitute one busy period after which there is some empty time.

And then there is a busy period and then there is an empty time. So, this x 1 r itself has busy period till this point from 0 till this red dot is the first busy period and likewise from the first renewal to this red dot is a busy period. So, this is a busy period. So, each of these x i r constitutes one busy period followed by some empty time and then there is another arrival to an empty system which constitutes a renewal.

We use stopping rule tool of stopping rule to prove that these are indeed renewals. These x i r are indeed renewals and we can also show that the number of arrivals during any of these renewal periods are also iid. So, essentially each of this busy period followed by some (()) (06:49) if you look at this renewal intervals they are all statistically identical copies of each other that is the key thing to note and this is very general.

This renewal aspects the renewals to empty systems constitute sorry arrivals to empty system constitute renewals is hold regardless of FCFS. It even holds for GGM you could have even multiple server. When the system finally empties when there is nobody in the system and there is a new arrival consecutive such instances are renewal intervals. So, this is what we will use for Little's theorem.

(Refer Slide Time: 07:36)



Avg occupancy = Avg arrival rate x Avg time spent in the system

Customers in the system at time $t \doteq L(t) = A(t) - D(t)$

(Time) Avg system occupancy $= \lim_{t \to \infty} \frac{1}{t} \int_0^t L(\tau) \cdot d\tau$ ← Does this limit exist?

Now what does Little's theorem say? As I said very colloquially I said that average occupancy is equal to average arrival rate average time spent in the system. Now what is the average occupancy? So, at any given time t so the number of customers in the system at time t = what? Let us say L of t is equal to what in this picture? It is the total number of arrivals minus total number of departures at that time.

Everybody who has come so far minus everybody who has left so far will be the number left in the system. This is some non negative integer this could be 0 or any integer. What is the average occupancy of the system. Now the question is are you talking about (()) (09:03) averages or time averages. You would expect that if the queue is ergodic it should not matter, but since we have studied time average reward so far we will look at time average occupancy.

So, average let me say time average system occupancy is equal to what this guy limit t tends to infinity 1 over t integral 0 to t L tau d tau. Now the question is does it even exist? I have just written out some limit t tending to infinity 1 over t integral 0 to t L tau d tau. There is no reason to believe that this limit exist. It is not clear. See what you can argue is this L tau is a legitimate reward function if you look at the renewal process defined by x 1 r, x 2 r and so on.

These renewals that entry arrivals to empty systems if you take these purple intervals in my picture x 1 r, x 1 r as the underlying renewal process. We can argue that this L of tau is a legitimate reward function for this renewal process namely at any given time L of tau depends only on the x random variables and v random variables in that busy period whatever that renewal period.

Therefore, if you take L tau a reward function for the renewal process defined by x 1 r, x 2 r and so on then by renewal reward theorem this time average will exist and in fact we know what it should be equal to almost surely should be the expected reward over a renewal interval over expected renewal duration and the expected renewal duration is simply expected x 1 r or expected x i r in general.

So, this is the approach we will use we will use renewal reward. So, what is if you take some time tau. Let us say I take some time tau out here what is this integral sorry I should say it should take some time t I guess. Let us say it take some time t what is this integral 0 to t L of tau d tau.

See L tau is simply A tau minus d tau. So, in the integral from 0 to t will simply be the area between the A and the d curves. So, this integral that I have circled in green will simply be that bit. Till t you have to integrate out and get that area and divide by t. Now, it is actually easy to see what this area is within a renewal interval. So, if you look at the sum of these areas in green so this is all a jump of 1 each time the jump is 1.

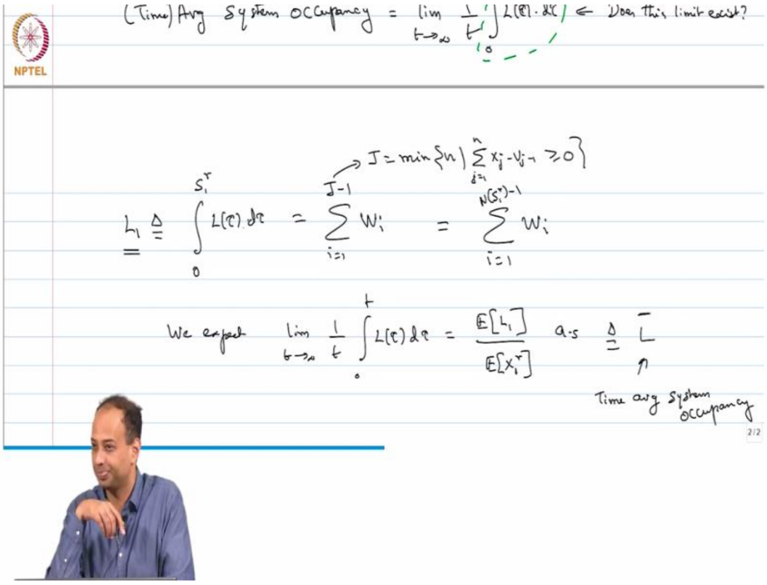
Now what is this width for example. Remember, this is the entry time of the x 1 customer first arrival after 0 and that person exits here. So, this is simply what it is not the waiting time waiting time it is not w q that we considered yesterday is the w. So, this guy will be w 1. Likewise if you look at this width, this width will be so this is where the x 2 customers enters and this is where he leaves.

So, this will be w 2 the total time spent in the system by that x 2 customer the second customer after one. Now please keep in mind that this w 1, w 2 are different from the w 1 q, w 2 q considered yesterday and when we are talking about the average time spent in the system we are talking about not w q but w is that clear. So, to get back to our question so of course this v 0 is equal to the time spent in the system by the 0th arrival.

So, what is this area under this green what is this green area. The width is the w 1 and the height is 1. So, the total area within an renewal interval is some of the waiting times not the waiting time some of the w the system times. So, there you see a relationship directly. So, the integral 0 to t to L of tau which is the L of tau is the system occupancy if you integrate that till some time it looks like you are adding the waiting time of all the customers you have seen so far.

And (()) (15:07) the fundamental relationship and this relationship will hold it really has nothing to do with first come first serve or last come first serve. This is the fundamental relationship the picture will change for some other service time, but this relationship will hold.

(Refer Slide Time: 15:24)



(Time) Avg System occupancy $= \lim_{t\to\infty} \frac{1}{t}\int_0^t L(\tau)\cdot d\tau$ ← Does this limit exist?

$$J = \min\{n \mid \sum_{j=1}^{n} x_j - v_{j-1} \geq 0\}$$

$$\underline{L_1} \triangleq \int_0^{S_1^r} L(\tau)\,d\tau = \sum_{i=1}^{J-1} W_i = \sum_{i=1}^{N(S_1^r)-1} W_i$$

We expect $\lim_{t\to\infty} \frac{1}{t}\int_0^t L(\tau)\,d\tau = \frac{E[L_1]}{E[X_1^r]}$ a.s $\triangleq \bar{L}$

Time avg System occupancy

So, let us look at this little more closely. So, let us say you look at this what is integral 0 to this instance is let us say S 1 r and this is S 2 r these are the renewal epochs of this purples process. Now what is integral 0 to S 1 r of L tau d tau. It is simply the sum of W i. See J is the index of the first renewal arrival to an empty system. So, it is really i = 1 to J – 1 where J is from yesterday where J is the you have to recall from yesterday.

Minimum n such that sum over j = 1 to n x j – v j – 1 greater than or equal to 0 is that right. This is what is this equal to? You are just integrating the reward over one renewal interval. So, this is like your r n so to speak r 1 this is we will call this L 1.Similarly, L n will be integral over the nth from S n – 1 r to S n r or whatever, but they are all iid. The issue is that L n define like so this is the definition for L 1 for the first renewal duration.

But you can define an L n similarly and these L n will be iid and that is the key issue. If you have that if these L n are iid then you will actually have Little’s theorem that is the point. Actually we can write this in a slightly different notation we have write this as sum over i = 1 to let us get this right N of S 1 r – 1 of w i. This may look like complicated notation, but it is really not that complicated.

So, S 1 r is this guy this so N of S 1 r is the index of the first arrival to a empty system N of S 1 r minus 1 is the arrival before this. So, you are adding all the waiting times of the arrival just of all the arrivals till just before the arrival to an empty system. So, that is why we put a N S 1 r – 1 so N S 1 r is your J if you think about it is exactly your J. It is the index of the arrival which sees an empty system.

Likewise, you can define L n I think I made a small notational mistake if you just go back the first customer has waiting time w 1 in my notation is it can you confirm. So, then I made a small mistake so I should write w 2 over here and w 3 over here then I will be okay. So, w 1 includes your 0th customer that is what I am saying so then this will be okay thank you for pointing out.

Now we are okay I think I was off by one basically now we are okay. So, this is great so I would expect this if renewal reward works out for this big renewal process this purple renewal process we expect this limit t tends to infinity 1 over t integral 0 to t L tau d tau should be equal to expected can you tell me expected what L n or L 1 does not matter. Expected L 1 over expected renewal duration which is x i r which is I can just write x 1 r almost surely.

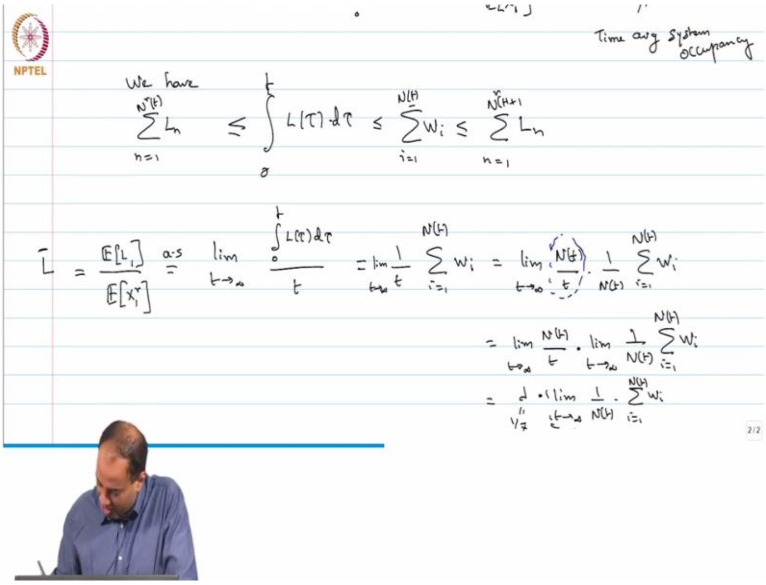
So, if you regard L of tau as a reward process for the big renewal process by big renewal process I mean this purple renewal process constituting arrivals which arrive to an empty system then we have this 1 over t integral 0 to t L tau d tau this limit as t tends to infinity. This time average exists almost surely and it is equal to this guy expected L 1 over expected x 1 r almost surely and if your expected x is greater than expected v meaning that the queuing system does not build up forever then this will be something finite.

So, great so this number exist. So, this is what I am going to call this number I am going to call my L bar. What is L bar? Time average system occupancy which is the time average number of customers in the system. So, this L bar exists now we want to prove that this L bar is equal to average arrival rate times averaging waiting time or system time in the system time spent in the system.

So, that is what we will prove using this relationship about this integral of L and the sum of w. So, if you look at this picture for any time t the integral from 0 to t is simply the sum of the w of the customers who have left so far plus maybe a little bit of people who have been waiting so far also which is this bit here. So, I will use a bounding approach as usual. So, what will I really do?

I will basically take this 1 over t integral 0 to t L tau d tau which is the area denoted in green here is lower bounded by the sum of all areas till the previous renewal interval and upper bounded by the sum of areas till the next renewal interval that is the relationship I am going to use.

(Refer Slide Time: 23:05)



Time avg System occupancy

We have

$$\sum_{n=1}^{N^r(t)} L_n \leq \int_0^t L(\tau)\,d\tau \leq \sum_{i=1}^{N(t)} W_i \leq \sum_{n=1}^{N^r(t)+1} L_n$$

$$\bar{L} = \frac{\mathbb{E}[L_1]}{\mathbb{E}[X_1^r]} \overset{a.s}{=} \lim_{t\to\infty} \frac{\int_0^t L(\tau)\,d\tau}{t} = \lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^{N(t)} W_i = \lim_{t\to\infty} \frac{N(t)}{t} \cdot \frac{1}{N(t)} \sum_{i=1}^{N(t)} W_i$$

$$= \lim_{t\to\infty} \frac{N(t)}{t} \cdot \lim_{t\to\infty} \frac{1}{N(t)} \sum_{i=1}^{N(t)} W_i$$

$$= \underset{1/\bar{x}}{\overset{\shortparallel}{\lambda}} \cdot \lim_{t\to\infty} \frac{1}{N(t)} \cdot \sum_{i=1}^{N(t)} W_i$$

So, I am going to say that so this is from the picture. We have for the picture itself you can say so integral 0 to t L tau d tau is lower bounded by sum over n equal to 1 to N r of t of L n and it is upper bounded by. So, L 1 is this green area under the first renewal period. So, if I go from integral 0 to t L of tau it is lower bounded by the sum of areas of these w it is lower bounded by L 1.

So, in general if t is somewhere in some renewal interval you look sum all the green areas till before that renewal interval you will get the sum over L n. So, this N r of t is I am counting the big renewals now the purple renewal N of t is the renewal arrivals, N r of t is the underlying renewal process of these arrivals to empty systems. So, this looks like extremely complicated notation, but it is really just saying this picture.

If you just have this picture you will be okay and I am upper bounding it by so this area under green is upper bounded if I were to go to the next renewal interval that is really all I am saying. So, if you then take 1 over t again so I maybe off by 1 I think sum over w i w 1 is the waiting time of the first this might be N of t I think. I think this includes the N of t arrival. This N of t is the arrival process N r of t is the renewal process.

There are two renewal processes here so it may be a bit confusing there is a renewal process of arrivals and there is a bigger renewal process of arrival to empty system. I have r for the bigger renewal process so to speak, but this notation is really complicated, but the picture is very intuitive. So, if you look at 1 over t of this so if I take limit t tends to infinity integral 0 to t L tau d tau over t.

So, this will be so you can basically put 1 / t throughout in a previous equation. So, this will actually be equal to 1 over t sum over i is equal to 1 to N of t w i. Why is that? Because this is upper bounded by 1 over t of sum of areas till the next renewals and lower bounded by the corresponding sum of areas till the previous renewal and this average we know goes to this we know goes to this is almost surely equal to what?

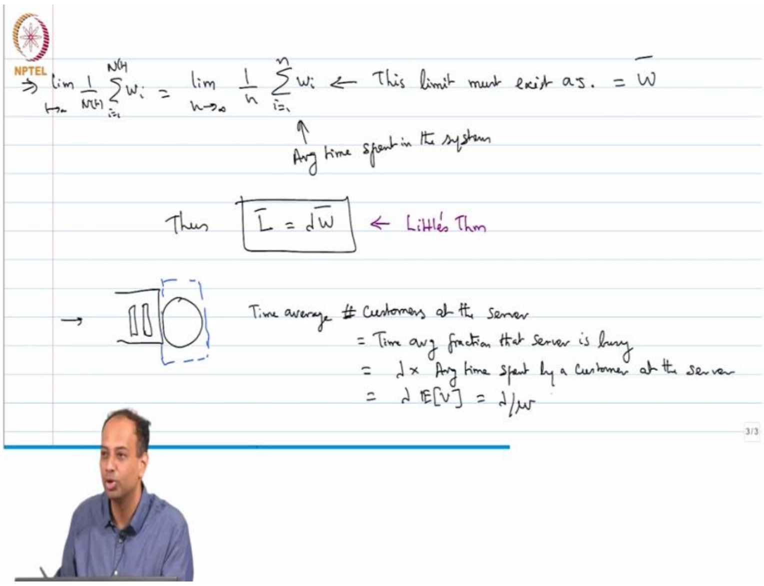
Expected L n over expected L 1 over expected I can just write x all these L i are identically distributed expected L 1 over expected x 1 r because I am using sandwiching because the sum over w i is sandwich between these sum of these L i. Here also there is a limit. So, this is of course your L bar which we have already defined. Now the question is what is this equal to here I can write it as limit t tends to infinity.

I can multiply divide by N of t over t times I over N of t w i. What have I done? I have just put N of t multiplied and divided by N of t. Now, this guy is very well known to us. What is it that equal to? N of t remember is the arrival renewal process so N of t over t goes to 1 over x bar which is lambda. So, I am going to write this as follows. So, this is equal to limit t tends to infinity N of t over t times limit t tends to infinity 1 over N of t that is equal to 1 to N of t w i.

This is equal to the lambda which is just 1 over x bar this lambda is just 1 over x bar times limit t tends to infinity 1 over N of t. Sum over i equals 1 to N of t w i. Now what is this equal to if you look at any ideas so 1 over N of t sum over i = 1 to N of t w i which is the you are adding all the system times of customers who have showed up till time t dividing by total number of customers who have showed up till time t.

What is that equal to? Average time spent in the system. Now, see there is no as such there is no reason for this limit to exist. If I just tell you that there is this limit t tending to infinity of 1 over N of t sum of i = 1 to N of t w i. So, we know that N of t goes to infinity.

(Refer Slide Time: 30:58)



$\Rightarrow \lim_{t\to\infty} \frac{1}{N(t)} \sum_{i=1}^{N(t)} w_i = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} w_i \leftarrow$ This limit must exist a.s. $= \overline{W}$

$\uparrow$ Avg time spent in the system

Thus $\boxed{\overline{L} = \lambda \overline{W}}$ $\leftarrow$ Little's Thm

Time average # Customers at the Server

= Time avg fraction that Server is busy

= $\lambda \times$ Avg time spent by a Customer at the Server

= $\lambda\, \mathbb{E}[V] = \lambda/\mu$

So, whatever is in this box is like this limit. If I just tell you limit n tending to infinity 1 over n sum over i w i i = 1 to n. This is just the average system time. There is no reason for this limit to actually exist as such, but because of this relationship between L bar. See L bar exist almost surely this integral 0 to t L of tau d tau over t goes to almost over limit L bar and of course limit t tends to infinity lambda N t over t goes to lambda almost surely.

Therefore, what remains must also (()) (31:44) almost surely. So, what does it show? So, this implies that limit t tends to infinity 1 to N of t w i which is simply limit N tending to infinity 1 over n i = 1 to N w i. This limit must exist and this is your average. So, this guy is your average time spent in the system let us say equal to w bar. So, what have you proven? So, we have proven L bar is equal to so again just going back and looking at this.

So, I will tell you how that happened. So, how would I make this equality that is the question. So, there is this inequality here. So 1 over t integral 0 to t L tau d tau should be less than or equal to 1 over t of all this, but then if I put 1 over t here in front of this so if I put 1 over t here and 1 over t here 1 over t throughout. Now send t to infinity this guy and this guy will go to the same limit because they will go to same limit L bar.

So, by sandwiching the limit will become equal is that clear. So, maybe I should do this so what I am saying you put 1 / t everywhere so this is 1 over t, 1 over t, 1 over t. This the upper bound and the lower bound will both go to expected L 1 over expected x 1 r. Therefore, limit t tending to infinity of the integral will be equal to limit t tend to infinity of sum of the w i and then we are saying that we are just multiplying, dividing by N of t.

N of t by t goes to lambda we know that strong law and then this beast that I have circled out here will have to go to a limit almost surely will have to why because left hand side is having an almost sure limit and N of t over t has almost sure limit. So, it cannot be that it has to be the case that 1 over N of t sum over i = 1 to N of t w i must have almost sure limit and that is your average system time average time spent in a system by a customer.

You are taking the sum total of all system times and dividing by total number of customers that is your average rating. See what is the crucial element in this? FCFS is not at all crucial FCFS is only for the picture. If it is a last come first serve your picture will be different, but the relationship will still hold. So, FCFS the service discipline is not at all relevant in fact even G / G / 1 is not very important you can do G / G / M also.

Even if you are M servers consecutive arrivals to an empty system will still be a renewal duration and we can still use the same sort of a mechanism to work it out even for G / G / M Little’s theorem will hold. See of course the value of L bar and w bar will be different, but when I say Little’s theorem holds I mean that L bar of that system will be equal to lambda times w bar of that system.

So, even if go from FCFS to some other discipline LCFS let us say the L bar for FCFS could be different from L bar for LCFS and w bar for LCFS will be different from w bar for FCFS, but nevertheless the equation L bar = lambda w bar will hold that is what I am saying. See when I say service discipline is not relevant I mean it is not critical I am not saying that everything will be the same.

I am saying that the relationship L bar = lambda w bar will still hold. Lambda bar will be different across different service disciplines generally L bar and w bar will be different across different service disciplines. So, this is Little’s theorem. We can also apply Little’s theorem to just the server. So, you have this G / G / 1 Queue with a server we have arrivals people queuing up and all that.

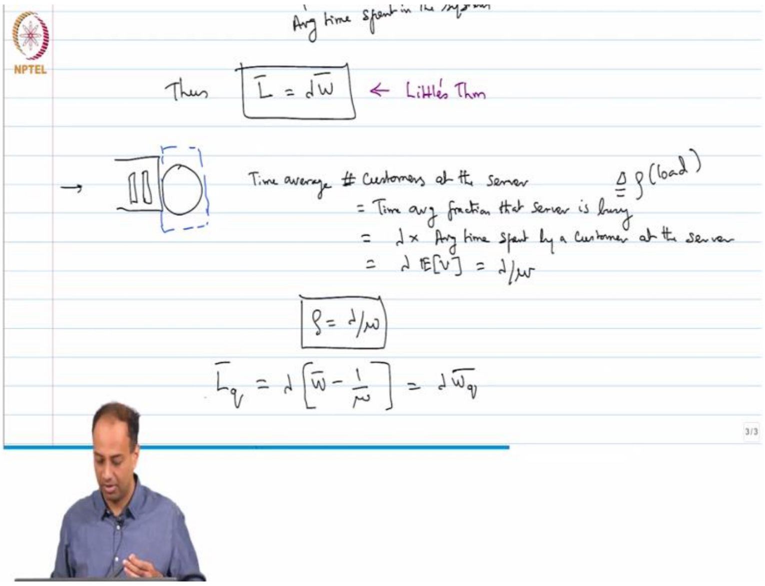
What you can do you can just put the server let us say inside a box. Now, if you just look at the server the server will see how many people can be in a server either 1 or 0. So, if the server is serving one person that person finishes and immediately the next person comes there is no gap. So, we will serve one person that person ejects another person comes it serve eject and then it become empty at some point.

So, now again we can argue just looking at the server consecutive arrivals to an empty server which means consecutive arrivals to an empty queue. Server is empty only if the queue is empty. So, consecutive arrivals that come to an empty server was again of course a renewals. So, we can use the same sort of argument to get a Little's theorem for the server same argument.

So, expected you can say so what is this not expected let us say time average number of customers at the server is also the time average fraction that the server is busy because time average number of customers at the server number of customers is just 1 or 0. So, time average of that will simply be the fraction of time that the server is busy it was just between 0 and 1 I mean it is either 1 or 0.

And the time average number of customers in server will simply be the time average fraction that the server is busy and this will be equal to what arrival rate lambda times average time spent by a customer at the server which is equal to what is the average time spent by a customer at the server 1 / average time spent by the customer. This is lambda times expected v which is simply lambda upon mu.

(Refer Slide Time: 41:13)



Avg time spent in the system

Thus $\boxed{\bar{L} = \lambda \bar{W}}$ ← Little's Thm

Time average # Customers at the Server $\triangleq \rho$ (load)

= Time avg fraction that Server is busy

= $\lambda \times$ Avg time spent by a Customer at the Server

= $\lambda\, \mathbb{E}[V] = \lambda/\mu$

$$\boxed{\rho = \lambda/\mu}$$

$$\bar{L}_q = \lambda\left[\bar{W} - \frac{1}{\mu}\right] = \lambda \bar{W}_q$$

So, the fraction of time that a server is busy is simply lambda over mu and this fraction of the time that the server is busy is denoted by rho this rho is called the load on the system. Load simply denotes what fraction of my time is my system busy. System busy means server busy. So, we get rho = lambda over mu. It is the arrival rate over service rate. This is another very fundamental relationship in queuing systems.

That the fraction of the time that the system is busy the server is busy is equal to average arrival rate over average service rate. This is again not just true for G / G / 1 FCFS it is true pretty much across the board. All you need is arrivals to an empty system must be renewals that is all that you need then this is true. Then you can just subtract so now if you look at L = lambda w bar what is L bar – rho?

See L bar is the total number of customers in the system rho is the fraction of customers see L bar is the average number of customers in the system. Rho is the average number of customers at the server so L bar – rho will be average number of customers who are waiting at the queue. So, you can subtract these two and get L q bar = what lambda is the same lambda times w bar – 1 over mu.

What is w bar – 1 over mu? W bar is a total system time by 1 over average total system time 1 over mu is the server time. So, this is equal to lambda w q bar. So, if you just look at the buffer alone or if you just look at the server alone some Little's theorem holds for them separately also. So, this is very powerful so even if you are given some complicated queuing system.

You can just look at some part of it and apply Little’s theorem for that part. Again, if you are given multiple queues as long as there is no correlation between arrivals and all that then you can still apply Little’s theorem. It is a very general and powerful. I will stop here.