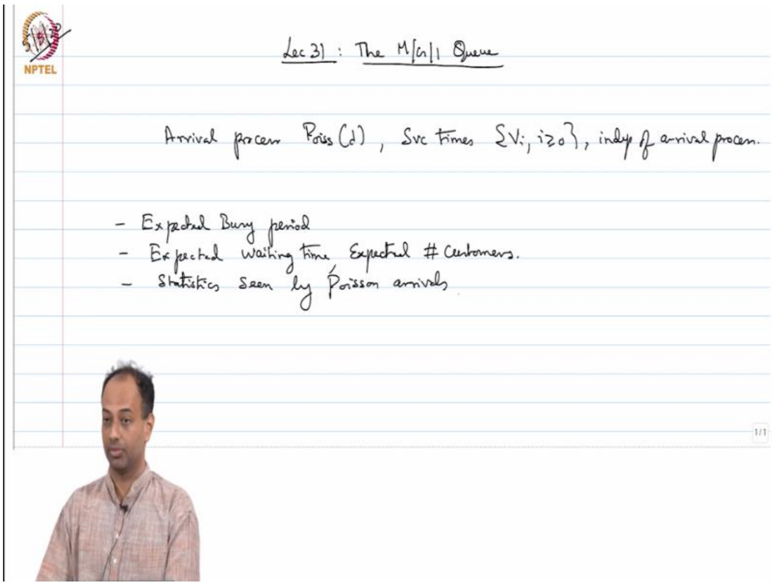
Stochastic Modeling and the Theory of Queues
Prof. Krishna Jagannathan
Department of Electrical Engineering
Indian Institute of Technology – Madras

Lecture –37
M / G / 1 Queue

(Refer Slide Time: 00:14)



Lec 31 : The M/G/1 Queue

Arrival process Pois(λ), Svc times $\{V_i, i \geq 0\}$, indep of arrival process.

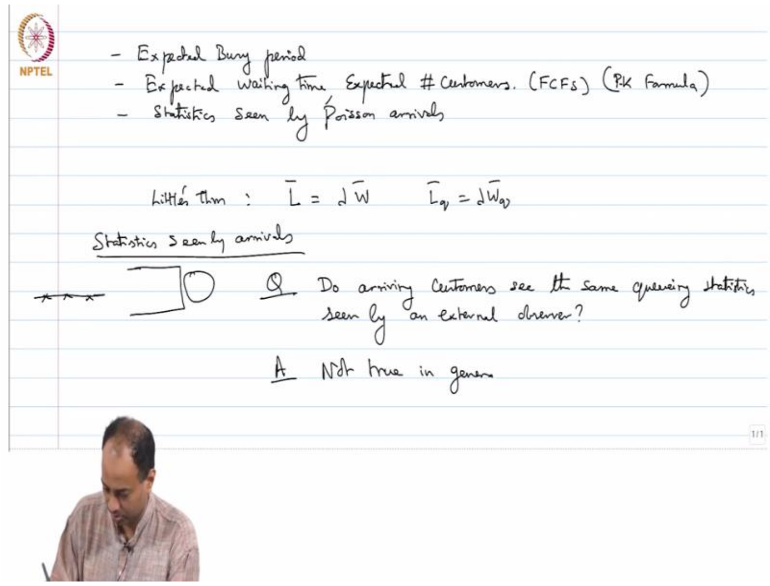
- Expected Busy period
- Expected waiting time Expected # Customers.
- Statistics seen by Poisson arrivals

Welcome back. Good morning. Today, you will study the M / G / 1 Queue. Yesterday we studied the G / G / 1 Queue. So, G / G / 1 Queue has renewal arrivals and iid service times with arrival process and service times being independent. The M / G / 1 Queue is a special G / G / 1 Queue where the arrival process is a Poisson of (()) (00:43) lambda service times as before some v i independent of arrival process.

So, the plan is for this M / G / 1 Queue we can derive some closed form expressions for parameters of interest such as the expected duration of the busy period, expected waiting time, expected number of customers in the system and so on. So, the agenda is we will derive expressions for expected busy period is enough, expected waiting time and expected number of customers in the system.

And we will also derive an important property about the statistics seen by Poisson arrivals. Actually we will need this path the third bullet to calculate the expected waiting time.

(Refer Slide Time: 02:19)



- Expected Busy period
- Expected waiting time Expected # Customers. (FCFS) (PK Formula)
- Statistics Seen by Poisson arrivals

Little's Thm : $\bar{L} = \lambda \bar{W}$ $\quad \bar{L}_q = \lambda \bar{W}_q$

Statistics Seen by arrivals

Q. Do arriving Customers see the same queueing statistics seen by an external observer?

A. Not true in genera

The second bullet about expected waiting time and expected customers we will do only for FCFS. So, for the expected waiting time there is a famous formula called the PK formula which we will derive. PK formula gives the expected waiting time in an FCFS M / G / 1 Queue. Now so in a G / G / 1 Queue we already know general we know that from Little's we know that L bar = lambda W bar for any queueing system which renews at arrivals to empty system satisfies this property and similarly L q bar = lambda W q bar.

The lambda times the average waiting time is equal to average number of customers waiting in queue. So, this is true very generally however for G / G / M or G / G / 1 Queues we know that this relationship holds, but neither L bar nor W bar is easy to characterize explicitly. We just know that they are related like this. For an M / G / 1 Queue it turns out that you can explicitly calculate W bar and W q bar.

And therefore you can apply Little to get L bar or L q bar. So, Little remains true, but neither of these average occupancy nor average waiting time is easy to calculate that is the issue with G / G / 1 M / G / 1 you can get closed form expressions. Now, what is it that so special about an M / G / 1 Queue which helps us calculate the quantities such as L bar and W bar and all that.

So, this is the path about statistics seen by Poisson arrivals. So, if you take some general queueing system let us say single server queueing system. Let us say G / G / 1 Queue and I have some let us say renewal arrival process to the queue. Now the question so this is the topic on statistic seen by arrivals. Question do arriving customers see the same queueing statistics seen by an external observer.

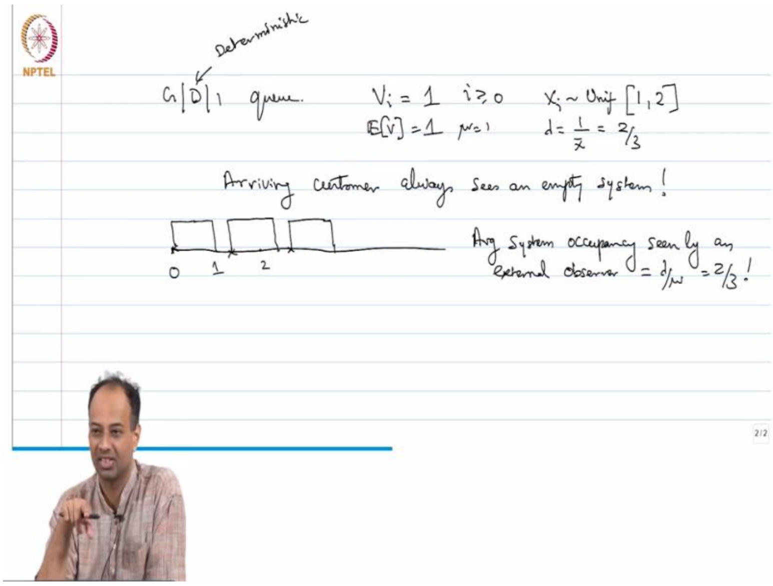
So, there is some queue that is running let us say G / G / 1 Queue that is running. As an external observer I am watching the queue evolve then I am able to calculate some average system occupancy L bar. I look at 1 over t integral 0 to t L tau d tau like yesterday or I calculate the average waiting time or average system time or whatever I want and I note it down I am external observer.

So, let us say I know L bar just G / G / 1 Queue has been running for a very long time I am able to calculate this time average or on ensemble average for that matter. For nice queueing systems these two are equal. So, let us say I am able to calculate L bar. Now the question is does an incoming customer to the queue. No I am no longer external if I am a customer who is coming to the queue can I expect to see the same kind of a statistics.

Meaning that do I face the same average delay or the same average number of customers in the system as an external observer that is the question I am asking. Is the question clear? So, is the statistics seen by an external observer which is just L bar or whatever or W bar same as the statistic seen by an incoming customer you are no longer extrinsic pseudo system you are coming in as an arrival as a part of the renewal process which is arriving to the queue.

A priori it is not clear right because I am a customer coming to the queue. I am a part of the renewal process that is driving the system and there is no a priori reason to straight away believe that the statistics seen by an incoming arrival is the same as that seen by an external observer. So, generally answer is not true. I will show you an example very, very simple example you can make a (()) (07:24) example.

(Refer Slide Time: 07:25)



Deterministic

$G|D|1$ queue. $V_i = 1 \quad i \geq 0$ $\quad X_i \sim \text{Unif}[1,2]$

$E[V] = 1 \quad \mu = 1 \qquad \lambda = \frac{1}{\bar{x}} = 2/3$

Arriving customer always sees an empty system!

Avg system occupancy seen by an external observer $= \lambda/\mu = 2/3$!

Let us say you have a G / D / 1 queue meaning what is D means deterministic the service times are so I am going to take the service time V i = 1 for all N. It is always each customer takes one unit of time to get served and let us say my arrival process I have to calculate let us say I have to characterize. So, let us say I am going to take x i’s are uniform in 0 to 1. So, each inter arrival time is a uniform random variable.

Meaning that so each time I pick a number between 0 and 1 uniformly at random and that will be my x i. So, all inter arrival times will lie between 0 and 1, but service times are deterministic. So, in the situation so expected V which is clearly 1 what is lambda? Lambda is we make so I think this will work for me. So, now I have lambda = 2 upon 3 if I do this what happens?

So, I have arrivals so I think it is uniform the inter arrival times are uniform in 1 to 2. So, in this case each customer takes one unit of time to get served and inter arrival times are uniform between 1 and 2. So, now what does an incoming customer see? See let us say I am an incoming customer. So, I will come in at some point when was the previous arrival? More than one unit of time ago.

So, if I come in at some point the previous arrival would have been at between one and two units of time before and that customer would have surely left why because in the service there is only one. So, whenever a customer coming in the previous customer would have arrived let us say time 1.2 seconds ago or whatever and that person would have completed and left by the time I come in.

So an arriving customer always sees an empty system. So, if you just plot this at 0 there is an arrival let us say we always take an arrival at 0. This guy will leave after so I am just plotting busy periods one unit of time. The next arrival will happen between one and two units of time let us say it happened here. This guy will take another one unit of time, but the next arrival after that will take another at least one unit of time after this you see what I mean and so on.

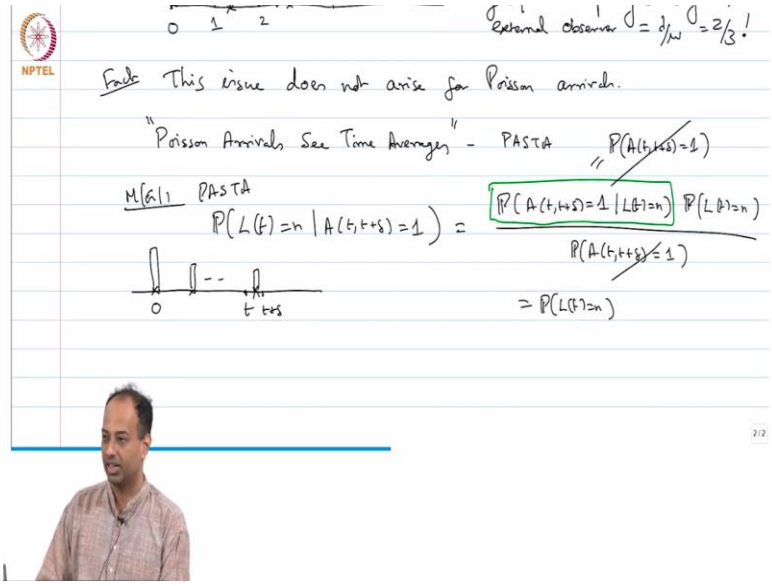
So, an incoming customer never waits. So, the average number of customers seen by an incoming arrival in the system. So, the average number of customers in the system seen by an incoming arrival is what 0. I never see anybody I think the system is always empty, but is the system always empty not at all, but average system occupancy is what? Average system occupancy seen by an external observer is lambda over mu that is the load on the system is not it.

So, the duty cycle of whatever I have drawn is 2 / 3. So, that is the surprising result. So, what is the moral of the story even in a very simple system I just cooked up an example right here. The statistics seen by an incoming arrival can be widely different from L bar here is lambda / mu it is the average number of customers and the system is also the average number of there is only customer in any given time.

So it is the average number of customers in the server which is lambda / mu. So, if you take any G / G / 1 Queue. So, this is the issue the statistics seen by an incoming arrival need not match what you calculate as L bar. I just showed you an example where L bar is not the same as this statistic seen by an incoming arrival so which is what makes characterizing the G / G / 1 Queue L bar or W bar explicitly difficult, but for an M / G / 1 Queue this is not the case.

So, the moment you have Poisson arrivals to a queueing system you will always have the Poisson arrivals see the same statistics as an external observer. So, Poisson arrivals to a queueing system always see average behave.

(Refer Slide Time: 13:54)



external observer $\rho = \lambda/\mu$ $\rho = 2/3$!

Fact This issue does not arise for Poisson arrivals.

"Poisson Arrivals See Time Averages" - PASTA

M/G/1 PASTA

$$\mathbb{P}(L(t)=n \mid A(t,t+\delta)=1) = \frac{\overset{\cancel{\mathbb{P}(A(t,t+\delta)=1)}}{\boxed{\mathbb{P}(A(t,t+\delta)=1 \mid L(t)=n)}}\ \mathbb{P}(L(t)=n)}{\cancel{\mathbb{P}(A(t,t+\delta)=1)}}$$

$$= \mathbb{P}(L(t)=n)$$

So, the issue is this fact this issue does not arise for Poisson arrivals. In fact the property known as Poisson arrivals always see time averages and this has a appetizing short form PASTA Poisson arrival see time averages or ensemble averages it does not matter. So, the issue is that Poisson arrivals always see statistics seen by an external person, person who is external to the queue.

So, this is what it happens so in an M / G / 1 Queue you have Poisson arrivals and that helps you calculate the average waiting time and all that. So, let us say an M / G / 1 sort of a system where there is Poisson arrivals. So, you look at probability that I want to look at this that L of t = n L of t is a total number of customers in the system at time t. So, I am looking at the probability that L of t = n.

Given that there was an arrival A, t, t plus delta is equal to 1. So, what am I saying so I am looking at so there is a queueing system that is evolving. I am looking at some particular time t + delta and I am telling you that there is an arrival in that small interval t, t + delta. I am going to look at this statistics given that there was an arrival at t, t + delta what is the probability that there were n customers in the system.

So, the number of customers in the system L of t = n will be because of all the these arrivals would have come with their own these are Poisson arrivals coming in with their own file sizes and the number of customers in the system at time t depends on all the arrivals and the file sizes of the customers who arrive before t. So, I am going to calculate this probably of L of t = n.

Given that I have a Poisson arrival coming in at the time t, t + delta. If I show that this is equal to the probability that L of t = n unconditionally then I have shown this property PASTA meaning that given that there is an arrival coming in now the probability that there are n customers in the system is the same as the probability that there are n customers in the size system at time t regardless of whether there is an arrival or not.

So, this is like probability of a given (()) (17:11) so I can write this as probability of A, t, t + delta = 1 given L t = n probability of equal to n over = 1 A t is the arrival process in the Poisson process A t is n t + 1 in all other rotations now this is great. So, now look at this term let me look at this term. What is the probability that I have an arrival in t, t + delta given that there are n customers in the system at time t.

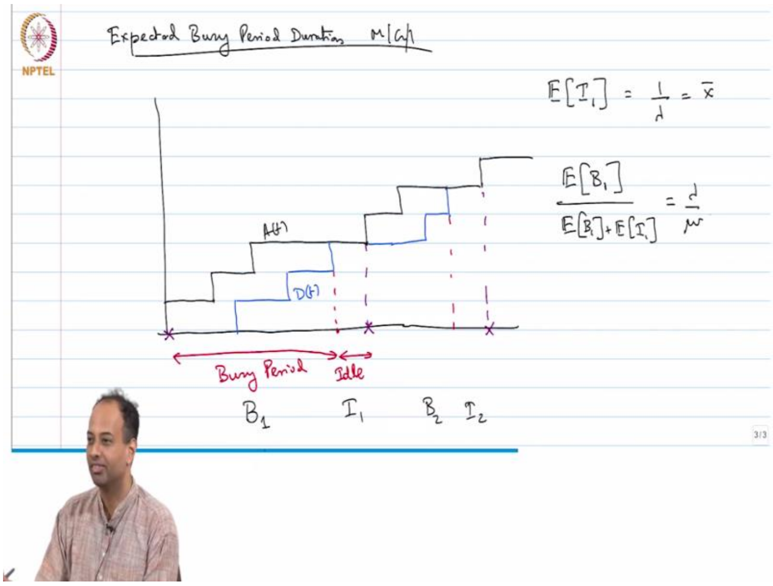
Now this L t = n the random variable L t is only a function of all the previous arrival instances and their file sizes. So, L t = n is actually independent of is the A t, t + delta. After all these Poisson arrivals has independent increment property. So, whether you have an arrival at t, t + delta or not is independent of all the previous arrival instance and by the construction of M / G / 1 Queue the arrival process itself is independent of V i the service time and L of t is only a function of previous arrivals their arrival instances and their V i.

So, what we can say is that the event of having an arrival in t, t + delta is independent of how many customers you have in the system at time t. So, what is the moral therefore? This conditional probability just becomes equal to so these are independent so this just becomes correct it is like the unconditional probability so that cancels with that. So, you get is this dependent on FCFS or any such thing.

It is not dependent I would have never used FCFS anywhere I have just used independence between I will use the independent increment property and the fact that V i’s are independent of the arrival times. So, this is what PASTA property really is. Poisson arrival see time average behavior. So, it is a sort of it is not just the Poisson arrival see time average behavior. They see the statistics seen by an external it is not only they see the averages are the same the statistics entire statistics are the same.

So, that is a key property this is true in great generality I mean this does not depend on single server, this does not depend on FCFS or any such thing. Now using this PASTA property we can calculate the expected waiting time and all that.

(Refer Slide Time: 20:39)



Expected Busy Period Duration M/G/1

$$\mathbb{E}[I_1] = \frac{1}{\lambda} = \bar{x}$$

$$\frac{\mathbb{E}[B_1]}{\mathbb{E}[B_1]+\mathbb{E}[I_1]} = \frac{\lambda}{\mu}$$

So, first let us calculate expected busy period duration or M / G / 1. So, you go back to this picture from yesterday you had plotted this remember this sort of arrival curve this is D of t and this is A of t and then there is another busy period and so on. So, we said that in a G / G / 1 Queue in general we had this is a renewal duration. So, when you have next arrival to the empty system you have this guy is again is a renewal duration.

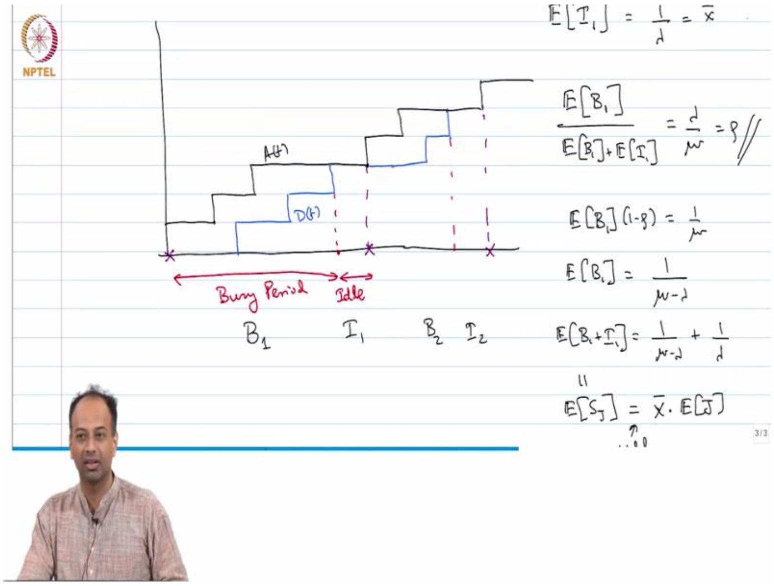
I want to know what is the expected duration of a busy period? So, what is my first busy period? So, a typical busy period is like this, this is a busy period and this is a idle period and I want to calculate the expected duration of these busy periods and idle periods. Now how do I do that is the question? Now this expected duration of the idle period is easy to calculate. So, here this is my system emptied at this point.

And then I am waiting for what to come in a Poisson arrival to come in. The Poisson arrivals are memory less. So, once my system empties how long does it take for Poisson arrival to come? It is an exponential with parameter lambda. So, this idle period is easy to calculate. So, let me calculate this, so let me call this busy period B 1 and this idle period I 1 and then there will be like this will be B 2 and I 2 and so on.

Now I have expected I 1 which is the same as any of the expected idle period is what 1 over lambda which is basically just X bar arrival process expected inter arrival duration. Now, do I know see I also know for an M / G / 1 Queue what is the fraction of time actually for any G / G / 1 Queue I know this what is the fraction of the time the system is busy. I did Little's theorem for the server alone.

And said the fraction of the time the server is busy is lambda / mu. So, I know this what is the fraction of time the server is busy expected B 1 over I am just taking the first we are all iid. This should be equal to lambda over mu. From this I can get expectation of B.

(Refer Slide Time: 25:24)



$$\mathbb{E}[I_1] = \frac{1}{\lambda} = \bar{x}$$

$$\frac{\mathbb{E}[B_1]}{\mathbb{E}[B_1]+\mathbb{E}[I_1]} = \frac{\lambda}{\mu} = \rho //$$

$$\mathbb{E}[B_1](1-\rho) = \frac{1}{\mu}$$

$$\mathbb{E}[B_1] = \frac{1}{\mu-\lambda}$$

$$\mathbb{E}[B_1+I_1] = \frac{1}{\mu-\lambda} + \frac{1}{\lambda}$$

$$\parallel$$

$$\mathbb{E}[S_J] = \bar{X} \cdot \mathbb{E}[J]$$

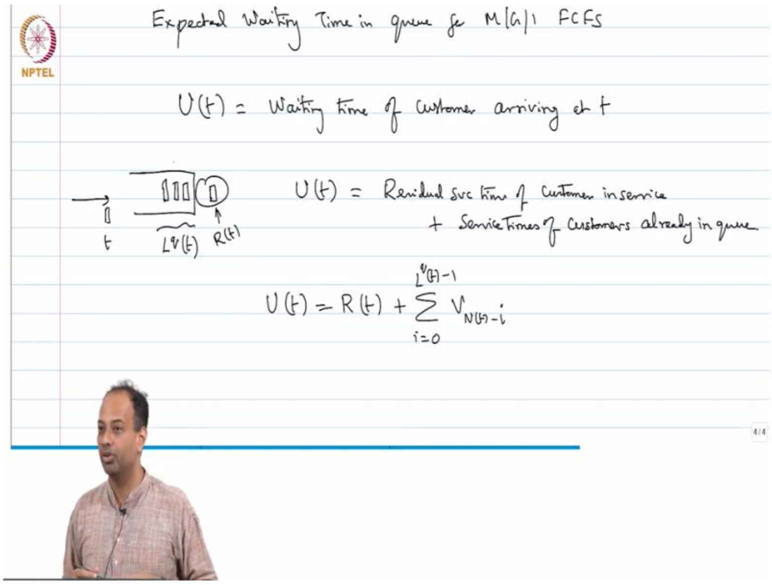
So you can rearrange this guy can you do this for me please what do you get expectation of (()) (25:28) I get expected B 1 times this is just rho right 1 over mu. So, expected B is equal to 1 over mu – lambda I think and the expected renewal so expected B 1 + I 1 also I can calculate is simply 1 over mu – lambda + 1 over lambda this is the expected S J from yesterday.

This is equal to expected S J where you know S J is the sum of the x i's till the first renewal. So, from here by Wald this will be equal to X bar times expected J where J is the index of the first arrival which is an empty system. So, from here I can even calculate expected J so this implies expected J is equal to can you help me now 1 over this should be a dimensionless quantity obviously mu – lambda is it all there is nothing else.

This is the expected J so J is the first index which sees a empty system and expected J – 1 will be the expected number of arrivals during a busy period. So, the expected number of arrivals during a busy period will be expected J – 1. So, all this you can easily calculate for an M / G / 1 Queue. So, the key issue here was how did we manage this? This cannot be done for a G / G/ 1 Queue.

So, which of this is true for G / G / 1 Queue see this equation out here is true for a G / G / 1 Queue just that is just Little's theorem only server. This is not true for the G / G / 1 Queue the first equation is not true because when somebody left the time to the next arrival is not memory less. In some residual time that we cannot easily calculate so that is a complicated matter, but for an M / G / 1 Queue this issue does not arise. So, expected busy period duration can be easily calculated.

(Refer Slide Time: 28:40)



Expected Waiting Time in queue for M|G|1 FCFS

$U(t)$ = Waiting time of customer arriving at t

$U(t)$ = Residual svc time of customer in service + Service Times of customers already in queue

$$U(t) = R(t) + \sum_{i=0}^{L^q(t)-1} V_{N(t)-i}$$

Likewise, we can calculate expected waiting time in queue for M / G / 1 FCFS. See what I said earlier is true generally it has nothing to do with FCFS. It is true for FCFS, but it can be anything the picture I have drawn is for FCFS, but what we have really used is Little's theorem which is generally true and that under any service discipline the moment I complete system empties the time for the next arrival is exponential that is always true.

No matter what the service discipline is, but now I am going to calculate expected waiting time in a queue for M / G / 1. So, I am going to denote U of t as the waiting time of customer at t. Waiting time meaning time I wait before I get into service. So, I have this M / G / 1

Queue there is somebody at the server let us say and then there is a whole bunch of people who are waiting and then I am looking they have this arrival coming in at time t.

How long do I wait this customer wait before I get into service this customer gets into service. See this U of t consists of two parts. So, what all do I have to wait for so I come in at some time t let us say this U of t the waiting time before I get into service consists of two components. First of all when I come in there is already somebody at this service that person will have the completed service.

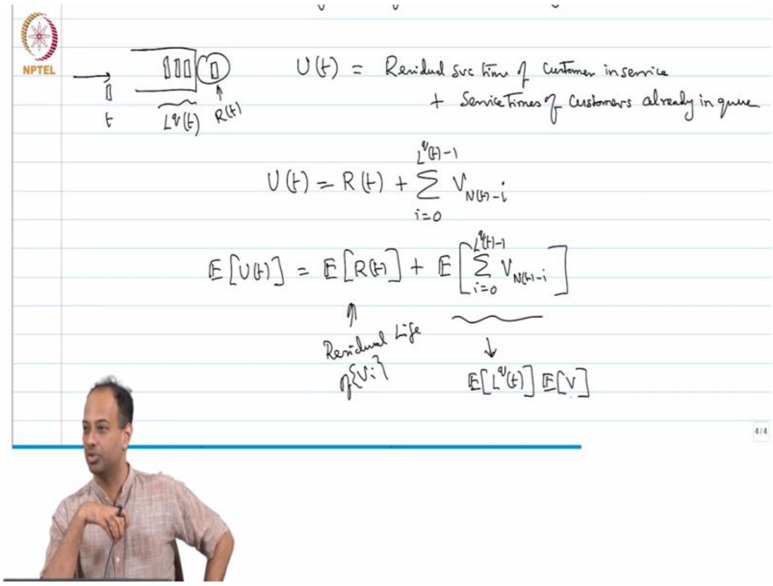
And then get out which is the residual time of the so this is like residual service time of customer in service plus then I have to wait for see I am doing FCFS. I have to wait for all the customers in front of me, but not in service to finish the service. So, I have to wait for the V i’s of all the guys in front of me, but not yet entered in service already in queue. So, let us say this number of customer let this be called L q of t this is the notation I am using.

L q is the number of customers who are waiting in the queue. So, I can write U of t = R of t this is the residual service time. This guy has another R of t amount of service left so he has half eaten so to speak when I come in the guy who is being served now some portion of this is eaten. So R of t is remaining, but the people in the service their service times are all iid they are the usual V’s and how many of this do I have to sum?

L q of those I have to sum L q of these Vi’s I have to sum. So I can write this as sum i = 0 to L q t – 1. V N t – i does that so V N t so the person coming in at time t so I am coming in at time t how many arrivals have come before me N of t arrivals have come. So, this person the person who is just ahead of me will have service time V N t see at time t N of t arrivals have occurred and V N t + V N t – 1 + dot, dot, dot how many L q t that many terms I have to sum. So, this is my relationship.

Now I want to calculate the average of U of t average U of t would be average waiting time of customer at time t. So, you can calculate either time average or on ensemble average and in this kind of a nice queueing system they will be equal.

(Refer Slide Time: 34:27)



U(t) = Residual svc time of customer in service
+ Service Times of customers already in queue

$$U(t) = R(t) + \sum_{i=0}^{L^q(t)-1} V_{N(t)-i}$$

$$E[U(t)] = E[R(t)] + E\left[\underbrace{\sum_{i=0}^{L^q(t)-1} V_{N(t)-i}}\right]$$

Residual Life of $\{V_i\}$

$E[L^q(t)]\,E[V]$

So, what can I do? I can write expected U of t = expected R of t + expected – i. So, in order to calculate expected U of t I have to calculate that these two expectations. What I really want is not just expected U of t, but the steady state as t tends to infinity. After a long time I want to know what the limit of expected U t is or time average I have taken ensemble average you can just take time average.

Now, which of these two is easier to calculate you think? This is residual time actually you know how to do both. This residual time has something to do with. So, there are all this customers which are getting service and leaving and when I come in that is the person who is getting served. So, the remaining service time of that person will be like a residual time of a renewal process you see what I mean where the V I's are the service times.

It is a little bit like calculating the residual life of a renewal process. So, this is somewhat like residual life of the V i process, but we know how to calculate. We draw all these isosceles triangles and this is actually even easier that is because actually why is that easier? Well, the V's are iid and the question is this are these random variables V is independent of the L q random variable.

See L q is the number of people waiting at time t the system at time t which is a function of all the arrivals that have happened so far and all the departures that have already taken place. Now this V i's are of course independent of all the arrivals and these V i's are also independent of the departures that have already taken place because those V i's are independent anyway of thee V i's.

So, you can argue that L q t is independent of the V i's of the customers in service. This will just boil down to expected L q t times expected V. So, we will continue this in next class, but you see where this is going you have one residual life term to manage which will have some isosceles triangle, pictures to be drawn and calculated and the next term has expected number of customers waiting in queue times their expected service time expected V.

Now what is L q of t? It is the number of customers in queue seen by a Poisson arrival by PASTA property this L q of t will be the average number of customers in the queue. See L q of t is the you see what I mean the customer incoming arrival sees time average statistics so this L q of t can be related to W q of t using lambda Little's theorem that is how you will proceed.

So, we will use PASTA property, we will use Little's theorem and we will use residual life calculation to get the waiting time in the M / G / 1 Queue. We will proceed in the next class.