


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

Lecture –38
M / G / 1 Queue & PK Formula

(Refer Slide Time: 00:13)

Lec 32: M/G/1 Queue & P-K Formula

Pois(λ) →  Service Times $\{V_i, i \geq 0\}$ iid & indep of arrival proc.
FCFS service discipline.


Want Expected waiting time in the queue experienced by an arriving customer in "steady-state".



Welcome back. So, we were discussing the M / G / 1 Queue and we were on our way to deriving the PK formula the Pollaczek–Khinchine formula which gives us the expected waiting time in an M / G / 1 Queue. So, just recall that you have this M / G / 1 Queue first come first served arrival process is Poisson of rate λ . Let us just say service times V_i are iid and independent of arrival process.

This is only for first come first served service discipline. Want to calculate expected waiting time in the queue experienced by an arriving customer in steady state. What do we mean by steady state? Steady state means that this system has been running for a very long time. So, we want to calculate expected waiting time spent in the queue by an arriving customer as t grows large.

(Refer Slide Time: 02:24)



Want Expected waiting time in the queue experienced by an arriving customer in steady-state.

$W^q(t)$ = Waiting time experienced by a customer arriving at time t .

$R(t)$ = Residual service time of the customer in service

$L^q(t) = \#$ Customers waiting in the queue at time t

$$W^q(t) = R(t) + \sum_{i=0}^{L^q(t)-1} V_i$$

Residual time Sum of service times of waiting customers

So, the way we did this so we had put down some equation if you remember so typically when arrival is coming there will be one customer who is getting served and a bunch of customers who are waiting in queue and a particular job arrives at a time t as a part of the Poisson process. So, let $W^q(t)$ they say I think I wrote it that $W^q(t)$ is the waiting time experienced by a customer arriving at time t .

And I denote it by $R(t)$ as the residual service time of the customer in service. So, customer arrives at time t that customer who would have to wait how long would he have to wait that customer has to first wait for there is somebody who is already at the server that person has to finish service so that residual time he has to wait and then he has to wait for the waiting time of all of these guys who are already in front.

So, remember we are doing first come first serve so you have to wait for all of this. So, the equation we wrote down in the previous lecture was that $W^q(t) = R(t) + \sum_{i=0}^{L^q(t)-1} V_i$ we can check this times then you have that much to wait is that correct. So, this is saying that you are waiting for the residual time plus $L^q(t)$ is what so let me just say what $L^q(t)$ is?

$L^q(t)$ is the notation for the number of customers waiting in queue at time t . So, this is the notation $L^q(t)$ (05:59) sum of service times of waiting customers.

(Refer Slide Time: 06:16)

Process a waiting time in the queue experienced by an arriving customer in steady state.

NPTEL

$W^q(t)$ = waiting time experienced by a customer arriving at time t .
 $R(t)$ = Residual service time of the customer in service
 $L^q(t)$ = # Customers waiting in the queue at time t

$$W^q(t) = \underbrace{R(t)}_{\text{Residual time}} + \sum_{i=0}^{L^q(t)-1} \underbrace{V_i}_{\text{Sum of service times of waiting customers}}$$

Want: $\lim_{t \rightarrow \infty} E[W^q(t)]$



So, now what do we want to find? We want to find expected W^q of t as t tends to infinity this is the steady state waiting time experienced by an incoming customer. So, in doing this we are going to invoke some key properties which you have studied before. One of them is PASTA property which is Poisson arrival see time average behavior. Remember this time t is not old time it is not some generic t .

It is a time at which what happens this particular customer is coming into the system. So, a customer arriving at time t and at time t the arriving customer sees a queue of $L^q(t)$ and sees a residual time of $R(t)$. Now, in general the residual time or the queue length seen by an incoming customer is not the same as typical time average behavior where the typical statistic seen by an external observer can be very different by the statistic seen by a arrival to the queue.

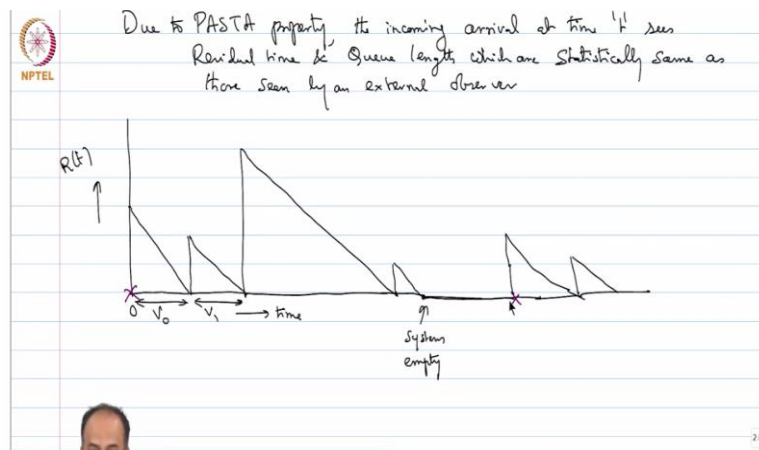
So, we gave an example if you remember last class we took a $G/D/1$ Queue and in general for a $G/G/1$ it is not at all the case that the residual time and the queue length seen by an incoming arrival need not be the same as the corresponding time average values. but for an $M/G/1$ Queue the arrivals are Poisson. So, what we will see is that these random variables at time t this $R(t)$ which is the residual time.

And this $L^q(t)$ will follow time average statistics and that is because of PASTA properties. The reason is that we are starting time at an instant when the first arrival the arrival comes at 0. The question is why are you starting at V_0 in that equation? If you remember we are considering time $t = 0$ time starts ticking when the first arrival comes whose file size is V_0 it

is just a matter of notation we have been following this notation, but it does not really matter it is okay.

Anyway what you want is you want PASTA here and then you want to be able to invoke these renewal reward type results for calculating R of t .

(Refer Slide Time: 09:21)



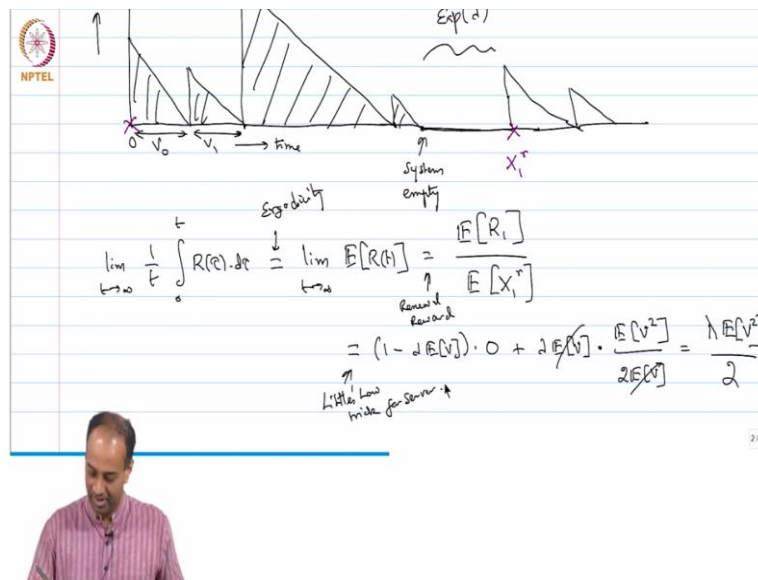
Due to PASTA the incoming arrival at time t sees residual time and queue length which are statistically same as those seen by an external observer. Now as regards R of t residual time how does the residual time behave? So, at time 0 a particular customer gets into service let us say the first this is I am going to plot what am I plotting I am plotting R of t against time. So, what is the time average behavior of R of t ?

R of t is the remaining service time of the job in service of the customer in service. So, the first job the customer who comes at time 0 has v_0 to begin with and then he is being eaten away at a constant rate; so you will get this sort of an isosceles triangle whose height is v_0 this is v_0 . I am just plotting the residual time left at the server and then there will be a v_1 suddenly jumps again starts getting eaten at a constant rate.

And then there is a v_2 it is probably quite large and there may be a small v_3 and so on and then let us say the system becomes empty it just might and then residual time is so let us say this system empties I am just drawing some typical R of t sample part then it is 0 for some time until the next arrival comes which will again have these triangles these have to be isosceles triangle.

This is the plot of the remaining. So, how much residual time does the job in service have left how long is it going to be at the server at time t and of course when there is nobody in service R of t is 0 which is what I have plotted here R of t is 0 out here. So, this is a renewal duration you know that so from 0 to the first arrival to a empty queue is a renewal duration and R of t constitutes a renewal reward process for this underlying renewal process.

(Refer Slide Time: 13:56)



So, you would expect the following, you would expect what is this? Limit t tends to infinity R of tau d tau which should ideally by ergodicity you would expect is the same as the ensemble average reward is equal to expected reward over one renewal interval over expected let us call this X 1 r the first renewal expected X 1 r or whatever let us say it is X 1 r the first renewal expected X 1 r.

So, you have to look at what is the total area under all these triangles divided by what is the expected renewal duration. You can calculate it like we did for you can calculate it directly also, but there is also a you can go ahead and do this. We already know I think from our homework problems what is the duration of these busy periods and idle periods after all what is the expected duration of the idle period here.

How long is this guy? This guy is equal to yeah so this is an exponential with parameter lambda exponential with mean 1 over lambda and what is the fraction of the time that the server is empty? Lambda times expected v or lambda over mu is the fraction of time it is

busy. So, we can easily find the expected time of these busy period. Expected idle periods are of course $1 / \lambda$.

So, we can find expected X easily and of course you can find the numerator expected R by just adding up the areas of these isosceles triangles. Another way to directly write this down is by again using Little's law. See with probability $1 - \lambda$ expectation of v the system is empty and with probability λ expectation of v there is somebody at the server. So, when there is nobody at the server R of t is 0.

When there is somebody at the server what is expected R of t ? So, I am trying to see I am trying to guess you can do this calculation expected R over expected X you can easily calculate by adding up all the isosceles triangle, but I am just trying to be telling you a different way of calculating this. So, after a long time I am looking at expected R of t after a long time.

So, this should be equal to with probability $1 - \lambda$ expectation of v which is my load on the system λ / μ and with probability λ expectation of v my expected R of t is what? It is expected v^2 second moment of the service time over 2 expected v . So what does this work out to be? This works out to λ expectation of v^2 over 2 .

So, I have made this your intuitive calculation, but you can check that this answer comes out even if you just calculate the sum of all these areas the expected reward over an renewal interval and divided by X you can calculate we know the busy period duration. You can please verify these. See I am making heavy use of PASTA property here because the expected residual time seen by an incoming arrival I am equating it to the time average residual time which I am again saying is equal to Ensemble average residual time.

This is by renewal reward you can calculate this ratio directly I am just making a Little's law trick here. Basically Little's law trick meaning I am looking at the Little's law trick for server. So, this is another way to calculate things. So, we know expected so we have this term so in this equation I want expected W of t which is this guy I have expected R of t as t tends to infinity. I just have to figure out what this second term is.

(Refer Slide Time: 20:14)

$$\begin{aligned} \bar{W}_q &= \lim_{t \rightarrow \infty} E[W_q(t)] = \lim_{t \rightarrow \infty} E[R(t)] + \lim_{t \rightarrow \infty} E\left[\sum_{i=0}^{L_q(t)-1} V_i\right] \\ &= \frac{\lambda E[V^2]}{2} + \lim_{t \rightarrow \infty} E[V] E[L_q(t)] \\ &\stackrel{\text{PAST}}{=} \frac{\lambda E[V^2]}{2} + E[V] \cdot \bar{L}_q \quad \leftarrow \text{Avg queue occupancy} \\ &\stackrel{\text{Little's}}{=} \frac{\lambda E[V^2]}{2} + E[V] \lambda \bar{W}_q \end{aligned}$$



So, I will write so go back to that equation so expected W_q of t limit this is what I want is equal to limit t tending to infinity expected R of t plus limit t tending to infinity expectation of this sum $i = 0$ to this is the notation $L_q(t) - 1 V_i$. So, the first term I know which I just calculated it is λ expectation of v square over 2 plus what is this? Now this is a sum of the v_i random variable how many of these v_i random variable am I summing?

I am summing another random number of these random variables. Now, L_q of t is the number of people waiting in the queue at time t . This L_q of t number of people waiting in the queue depending only on the past arrival time and the past departure times. So, it depends on the arrival times of customers who come so far and the departure times of customers who have already left.

In other words the number of people who are waiting in the system at this time t L_q of t is actually independent of the service times of the people who are not even entered service they have come to the system. This v_i 's are the service times of customers who are waiting because L_q of t is only a function of arrivals which have come in the past and the service times of the customers who have already left which are again independent of the v_i 's of the customers who are waiting in queue.

So, this is a random number of random variables being summed where this L_q of t the number of random variables you are summing is independent of the v_i 's. So, this is like this is not even Wald this is even simpler than Wald. I mean this is a trivial case of Wald if you like

where the number of terms you are summing is independent of the x_i 's or in this case v_i . The issue is that L_q is independent of v_i 's for $i = 0$ to L_q .

So, using that I can write this as limit t tends to infinity expected v times expected L_q because if you are summing a number of random variables where the number of random variables is independent of random variables of summing you get product of the expectation this is like a very simple version of Wald now even Wald is not necessary. So, now what is expected L_q ?

It is the steady state q occupancy seen by an incoming arrival. By PASTA L_q also has the same distribution as the time average queue occupancy. So, again take this remark here due to PASTA because of this remark expected L_q seen by an incoming Poisson arrival the same as the time average q occupancy. So, this can be written as $\lambda \bar{v}^2$ over 2 the first term will be same plus this is expected v which is times L_q .

What is now L_q bar this is by PASTA. See what am I saying expected L_q which is the expected occupancy seen by an incoming arrival is same as the time average queue occupancy that is PASTA. Now what is L_q bar? See what do you want to calculate the average waiting time in the queue. So, can I not write so let me call this W_q bar is the same as if you believe time averages and ensemble averages can be interchanged this is plus expectation of v times λW_q bar how is this this is by Little.

(Refer Slide Time: 26:54)

NPTEL

$$\stackrel{\text{PASTA}}{=} \frac{\lambda E[v^2]}{2} + E[v] \cdot \bar{L}_q \leftarrow \text{msg } v$$

$$\stackrel{\text{Little}}{=} \frac{\lambda E[v^2]}{2} + E[v] \lambda \bar{W}_q$$

$$\bar{W}_q = \frac{\lambda E[v^2]}{2(1-\lambda E[v])}$$

P-K formula
Pollaczek - Khinchin

only for FCFS
M/G/1

$$\bar{W} = \bar{W}_q + E[v] = \frac{\lambda E[v^2]}{2(1-\lambda E[v])} + E[v]$$

$$\bar{L} = \lambda \bar{W} \quad \bar{L}_q = \lambda \bar{W}_q$$


Which means W_q is equal to what can you help me with this? λ expectation of v^2 by twice $1 - \rho$ – I know the answer this is what it will come out to be you can rearrange it. So, this is the famous PK formula I wrote it down so confidently because this is the standard formula. You just rearrange the above equation you get this. This is called PK formula I am going to spell this incorrectly I think I think this is a Pollaczek–Khinchine.

This is the famous Pollaczek–Khinchine engine formula for the average waiting time than $M/G/1$ Queue. This is only for FCFS. The service discipline is different you cannot use this formula because we have very crucially used the fact that you are waiting for if you look at this picture we are waiting for the person in service and then waiting for everybody who has arrived before you.

This is directly using the FCFS nature of service otherwise this formula will not hold. So, the key intuition so what is the most striking thing about this formula? The expected waiting time in queue is proportional to the second moment of the service time. What is the intuition? The intuition is that see suppose see this v_i 's are iid so some are small and some are big the way I have drawn it for example I have drawn some big triangles and big triangles.

So, typically when you have a scenario like this with all this isosceles triangles you tell me this. This arrival process is the Poisson process which is uniform coming at a rate λ . So, if you draw the arrival process more arrivals are going to come here and very few arrivals are going to come during smaller service times. So, particularly large customer or large file is being served you will have a lot of customers coming.

And they will all be held up by this very large file, very large customer. See this is just like if you are showing up to a bus stop buses are a renewal process arriving at a bus stop according to a renewal process you are more likely to show up when the duration between the two bus arrivals is very long. It is a bit like that. So, typically the incoming Poisson arrival comes in when there is a large file in service. So, you can view this as this mice and elephants.

Typically, the files are of some average size, but when a particular Poisson arrival comes into a queue the file in service at the server the customer in service is an elephant. The residual time is proportional to the second moment of the service time. So, expected v^2 can be

very large compared to expected v . The second moment can be made arbitrary large for a given expected v that is possible.

So, in that sort of a scenario you will have this FCFS $M / G / 1$ Queue will have the customers coming in they face very large expected delays. Of course, for $M / M / 1$ Queue what happens the residual service time of the customer in service will also be exponential because the file sizes are exponential with some parameter μ . So, you can check that I mean you can verify PK formula for the specific case of $M / M / 1$ Queue and you will get a simple formula for $M / M / 1$ Queue.

So this is W_q bar which is the expected waiting time in queue and W queue which is just let us say W bar this is the average system time will be W_q bar plus how much that customer service time. So, this will just be again λ expected v square over twice 1 minus plus expected v and from here can you calculate L bar and L_q bar can also be calculated L bar will simply be λW bar by Little and L_q bar will simply be λW_q bar.

So, all this can be calculated for $M / G / 1$ Queue. So, we are assuming that expected v square is finite the second moment is finite.