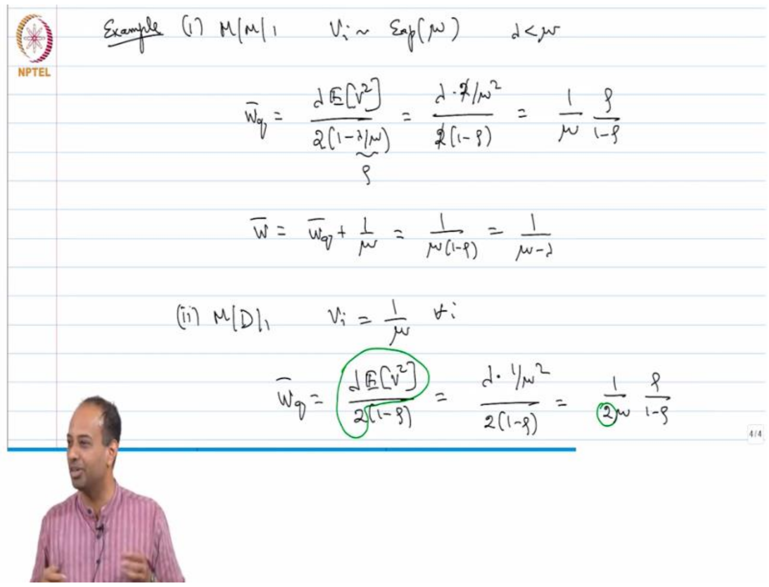
Stochastic Modeling and the Theory of Queues
Prof. Krishna Jagannathan
Department of Electrical Engineering
Indian Institute of Technology – Madras

Lecture –39
M / G / 1 Queue & PK Formula Continued

(Refer Slide Time: 00:16)



Example (i) M/M/1 $\quad V_i \sim \text{Exp}(\mu) \quad \lambda < \mu$

$$\bar{W}_q = \frac{\lambda \mathbb{E}[V^2]}{2(1-\underbrace{\lambda/\mu}_{\rho})} = \frac{\lambda \cdot \cancel{2}/\mu^2}{\cancel{2}(1-\rho)} = \frac{1}{\mu}\frac{\rho}{1-\rho}$$

$$\bar{W} = \bar{W}_q + \frac{1}{\mu} = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu-\lambda}$$

(ii) M/D/1 $\quad V_i = \frac{1}{\mu} \ \forall i$

$$\bar{W}_q = \boxed{\frac{\lambda \mathbb{E}[V^2]}{2(1-\rho)}} = \frac{\lambda \cdot 1/\mu^2}{2(1-\rho)} = \frac{1}{\boxed{2}\mu}\frac{\rho}{1-\rho}$$

So as specific examples if you have M / M / 1 where v i's are exponential with parameter mu. M / M / 1 means arrivals are Poisson, service times are exponential with parameter mu of course lambda is less than mu. In this case what happens? W q bar is lambda expected v square / twice 1 – lambda over mu. What is expected v square? Second moment of an exponential what is that equal to? 2 over mu square the second moment is 2 over mu square.

So, lambda times 2 over mu square over so this I can call rho lambda over mu is the load on the system 1 – rho. This 2 and 2 cancels that is equal to W q bar is 1 over mu times rho over 1 – rho and W bar is simply W q bar plus expected service time of that job which is 1 over mu which is equal to 1 over mu q – lambda. This is just algebra I am just plugging into PK formula.

So, we have M / D / 1 what happens? M / D / 1 means service times are equal to mu a constant. So, v i's are equal to mu for all i. So, in the first case the M / M / 1 case v i's are exponential with parameter mu in the second case I am considering M / D / 1 that is service times are deterministic so each service time is equal to mu actually I should say 1 / mu for all i.

So, in this case what happens W q bar is equal to same formula expected v square over twice 1 – rho. So, what is expected v square? V is just 1 over mu it is just a number. So v square is always 1 over mu square so you will get lambda times 1 over mu square over twice 1 – rho which is equal to 1 over 2 mu times W q bar is 1 over 2 mu times 1 – rho. So, how does the M / M / 1 and M / D / 1 queue compare, how do they compare?

So, if I have a given arrival rate lambda and a given expected service time which is 1 over mu let us say I hold lambda and mu constant, but in case 1 it is exponential service times in case 2 it is deterministic service times. So, my waiting time in the queue is a factor of two better when the service times are deterministic for the same service rate and arrival rate. Arrival process is the same Poisson process.

Service process is different one case is exponential other case is deterministic, but the rate is the same. So, the M / M / 1 Queue has twice the waiting time on an average compared to M / D / 1 Queue. Why is that intuitively? Intuitively what is happening? See in the M / M / 1 case the residual service time is also an exponential so the variance of the exponential is greater the variance of the M / D / 1 Queue is 0 there is no variance in play.

So, the residual time is basically only half as long in some sense so you will get this sort of a behavior. After all see the waiting time in the queue is simply lambda this quantity what is this lambda expected v square / 2 is what? The expected residual time over 1 – load. So, if we fix the load so the expected so what matters is the expected residual time. In their exponential case the residual time is what is the expected residual time?

See, when a job comes in the residual time of the customer and service is still exponential which mean 1 over mu whereas if you come in to an M / D / 1 system the person in service will have see typically it takes only 1 over mu to deterministically to serve a person. So, when you come in a Poisson arrival will see half of that half the job would have been eaten on an average only half will be pending.

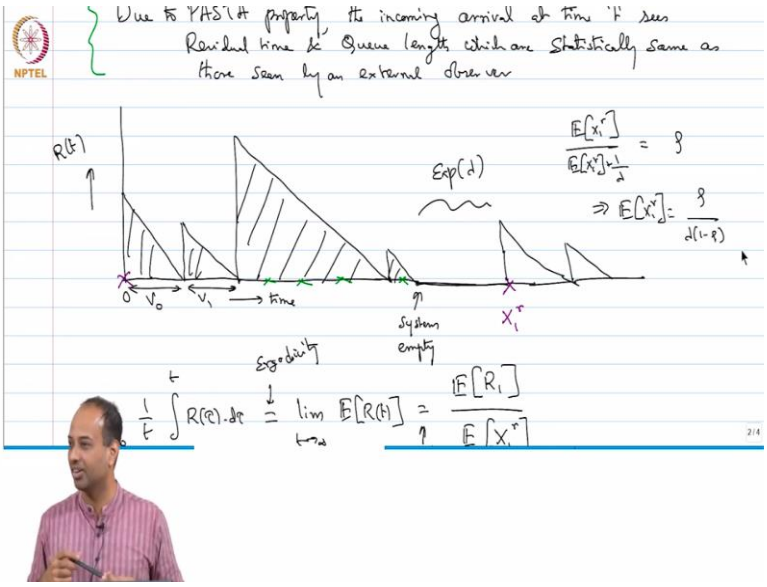
So, it will be the residual time will be half as small I mean half as big. So, you have half the waiting time. So that is the intuition so that is why you get a factor of 2 which is not there in the M / M / 1 Queue. See there are more extreme examples you can make up. You can keep the load the same and make expected v square very, very large. In fact you can make it unbounded you can even make it go to infinity.

In fact this time average renewal reward theory is still valid when expected v square is unbounded. So, you can have a system, you can have an M / G / 1 system which is stable in the sense that the queue I mean infinitely often you will have the queue renewing meaning that there will be renewals to the empty system infinitely often, but the expected waiting time in queue will become infinity because expected v square is you can make expected v square infinity.

You can make up distributions like this variance can be made infinity for a given expected v. So, you can make queuing delay arbitrarily large for a given load by just playing with the second moment. You are just making the elephant in service I mentioned can be made very, very large by taking a distribution which is like 1 over x cube or something like that you can just (()) (08:34) up a distribution where expected v square can be unbounded.

So, you can make expected waiting time in an M / G / 1 Queue can be made unbounded although the system is renewing infinitely often. So, the funny thing is that you will see expected renewal times will be finite, but the expected waiting time seen by an incoming arrival will be infinite.

(Refer Slide Time: 09:05)



Due to PASTA property, the incoming arrival at time 't' sees Residual time & Queue length which are Statistically same as those seen by an external observer

$$\frac{\mathbb{E}[X_1^r]}{\mathbb{E}[X_1^r]+\frac{1}{\lambda}} = \rho \Rightarrow \mathbb{E}[X_1^r] = \frac{\rho}{\lambda(1-\rho)}$$

$$\frac{1}{t}\int^{t} R(\tau).d\tau \overset{\text{Ergodicity}}{=} \lim_{t\to\infty} \mathbb{E}[R(t)] = \frac{\mathbb{E}[R_1]}{\mathbb{E}[X_1^r]}$$

What am I saying if you go back to this picture expected X 1 r is how much? Right here in this picture we know that expected X 1 r over expected X 1 r + 1 over lambda is equal to how much lambda this is just rho lambda expected v. So, from here what is expected X 1 r can you just tell me quickly? Is that all there is. So, it just now depend on see this is just lambda expected v over 1 over lambda times 1 over 1 – lambda expected v. So, this has no dependence on second moment.

It has only depended on expected v so you fix an expected v as 1 over mu you fix it you fix a rho. So, your expected renewal durations are finite, but the incoming job the incoming customer is seeing a unbounded expected waiting time because the incoming customer sees expected v square not just expected v an expected v square can be made very, very large for a given expected v for a given rho.

You can make it as large as you want. Is there a contradiction here? There is no contradiction because when you are seeing from outside you see the behavior of this renewal process renewing in finite expected time, but an incoming Poisson arrival sees an elephant in service whose residual time itself is expected v square lambda expected v square over 2 and that can be very, very large.

So, person who is observing from outside will see the system renewing infinitely often, but at given time the job in service is very huge in expectation. So the Poisson arrival which sees the elephant in service will wait for a very, very long time if expected v square is very large see what I mean. So, what we are saying is that the expected waiting time in queue can be much larger than the expected busy period duration.

You may think that the busy period is sum of all the waiting time, but that is not really true it is sum of the waiting times, but the person who is coming into the queue sees an exceptionally large job at the server and person who is seeing the entire system evolve over a long time is not seeing he is not pre disposed to seeing very large file sitting at the server you see what I mean.

It is a little bit like your bus stop example. So, let us say buses are going according to a renewal process in a bus stop. If I am just stopping at the bus stop and noting down all these inter bus arrival times I mean what I will note is typically arrival statistics, but if I were to show up to the system at a particular time the duration of that inter bus arrival will look like expected x square which will not be typical at all which we have already seen.

That is exactly what is happening here. So, there is nothing more mysterious happening in the M / G / 1 Queue than in the simple bus stop example. So, a person who is seeing the system evolve continuously we will see different statistics from a person who shows up to the system because that person is more likely to show up in the M / G / 1 case when the elephant is in service or in the bus stop case when the inter arrival duration is exceptionally wrong.

So, that is the key intuition I want you to get out of this PK formula, but this is only for first come first served because you are waiting behind for bunch of people who are stuck behind that elephant and that elephant in service is what is killing you so that is PK formula. We will continue tomorrow with ensemble average rewards.