Introduction to Queueing Theory

Prof. N. Selvaraju
Department of Mathematics
Indian Institute of Technology Guwahati, India

Lecture - 40
M/G/1 Queues: Waiting Times and Busy Period

Hi and hello, everyone. What we have been seeing in the last week was the semi-Markovian queues, and especially the $M/G/1$ queueing system or queueing model is what we have seen. What we have done is that first, we obtained the PK mean value formula directly without getting into the business of obtaining the stationary distribution or equilibrium distribution, or steady-state distribution. We directly obtained the mean value formula; then, we also obtained the system size distribution by looking at the number in the system at the departure epochs. And then, we showed that the departure point system size probabilities are exactly the same as the arbitrary point system size probabilities. And hence like that whatever distribution that we have obtained, of course, this is in terms of transforms only we gave because that involves the specification of service time distribution. Only when you give the service time distribution explicitly then, you can specify the distribution of the number in the system; you can obtain it from its transform. Till that time, we can give only the formula as to how one can do, and we have also set how one can do, and we have taken the simplest case of the service time being exponential. Then we showed that the transform or the generating function or the PGF of the system size distribution is exactly the same as the one that we have obtained for the case of an $M/M/1$ queue; that is what we showed it. And then the condition for the existence of that remains the same, $\rho < 1$. But it can also be proved using the result that we gave at the very beginning of the course when we talked about Markov chains, how one can view or one can see, or can prove the chain is positive recurrent. So, we gave one sufficient condition and using that sufficient condition, one can show that $\rho < 1$ is the necessary and sufficient condition for the existence of a steady-state.

So, if you are interested in looking at how exactly that thing falls, it is not a very difficult one, but due to lack of time, we are not going into the detail, but it can be shown so that we are assuming away. So, now, along similar lines as we have done for at $M/M/1$ queue, we would also want to look at the waiting times with respect to these $M/G/1$ queues. We already know there is a Little's law relationship that holds at the level of the first moment. L, which is the first moment of the number in the system, and W, which is the first moment of the system waiting time or the sojourn time, and these two are related through Little's law as $L = \lambda W$, that is what we have seen. Now, the natural question then one would enquire that is there any relationship that one can establish between the higher order moments, it may not be just the ordinary moments or the raw moments, it could be central moment factorial moment whatever it is, whether some moment because that if you get into one moment, you can get the other moments as well. So, it is one and the same. So, what is the relationship that you can establish? Or if there is any relationship between the higher order moments of these two quantities, meaning on the one side, you have the system sizes, and on the other side, you have the waiting times. Or, more generally, can we establish a relationship between these two distributions? It may be difficult to specify exactly what the relationship would be, but at least you can establish what the relationship between them is. It is not that precise, but it will be in terms of, as long as you do not specify the service time distribution completely; it can only be in a formula form. Once you specify that, it will become the specific relationship between the distributions or, if not in distribution level, maybe equivalently in terms of transforms.

And in fact, it will turn out that the relationship you are going to get here in our case of $M/G/1$ is basically the relationship in terms of transforms because, as we know, if you know the transform, then you also equivalently know the distribution. So, that is what the question is; is there a relationship between the moments, and is there a relationship between the distribution we can establish. So that, everything in a complete package kind of thing like you can get now, if you know one. That is what we are going to look at that.

To start with, note that the stationary distribution for the $M/G/1$ system can be written in terms of the waiting time CDF. We know that T is the random variable that we usually denote the system time, and $F_T(t)$ is the CDF of that. So, we can write the stationary distribution; we have already shown that $p_n = \pi_n$; π_n is the departure point probabilities, p_n is arbitrary time point probabilities. So, this p_n will be equal to π_n, is basically

$$p_n = \pi_n = \frac{1}{n!}\int_0^\infty (\lambda t)^n e^{-\lambda t} dF_T(t), \quad n \geq 1$$

How do we are writing this?

Imagine this is so because the system size under first come first serve will equal n at an arbitrary departure point if there have been n arrivals during the departures system wait; you think of a tagged customer who arrives to the system at that point of time. How long he will wait depend on how many customers are ahead of him in the queue. So, his system waiting time starts at the moment when he arrives to the system, then he waits sometime in the queue possibly or no wait as well. But he will go into the service at some point of time. Either immediately or after a delay in the queue, he will go to the service, then there is a service duration. So, from starting from his arrival until the service completion, what is his system waiting time? Now, what is the number of customers who will be left behind by this particular customer is the ones who arrived during his waiting time in the system. Because only that many numbers of customers he will leave behind, and that is what precisely this. He would leave behind n if there were n customer arrivals during his system waiting time. Suppose if I fix the system waiting time to be t, then in that interval of length t the number of arrivals of this Poisson process arrival, because remember this is M Poisson process arrival. So, it is basically distributed as a Poisson distribution with parameter λt, and hence you have $\frac{e^{-\lambda t}(\lambda t)^n}{n!}$ is what is the n arrivals would have been there during his system waiting time. Now, as you enumerate over all possible values of t, that is what you will get $\frac{1}{n!}\int_0^\infty (\lambda t)^n e^{-\lambda t} dF_T(t), \quad n \geq 1$. That is, for π_n, we see the probability that he leaves behind n, and that will be the same as the arbitrary time point of the system being in state n.

So, this is what you can observe easily, that the stationary distribution in terms of the waiting time CDF you can write it in this form. Now, what we can do, we can multiply this equation by z^n and sum over n; p_n expression then you will get $P(z)$, which is the probability generating function of the number in the system which is equal $\sum_{n=0}^{\infty} p_n z^n$. So, which is basically

$$P(z) = \sum_{n=0}^{\infty} p_n z^n = \int_0^\infty e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda t z)^n}{n!} dF_T(t) = \int_0^\infty e^{-\lambda t(1-z)} dF_T(t)$$

.

But now, what is $\int_0^{\infty} e^{-\lambda t(1-z)} dF_T(t)$ expression all about? This expression is also what you know now. $\int_0^{\infty} e^{-\lambda t(1-z)} dF_T(t)$ is nothing but the Laplace-Stieltjes transform of $F_T(t)$. So, let us call that as $F_T^*(s)$, star is we are using it for Laplace-Stieltjes transform, but with now the variable here is $\lambda(1-z)$. So, what we are obtaining is

$$\boxed{P(z) = F_T^*[\lambda(1-z)]}$$

.

So, now, you see that $P(z)$ equals the Laplace-Stieltjes transform of the system waiting time T with the variable as $\lambda(1-z)1$. So, these two are equal. This is the first relationship that you need to remember, that how the system's waiting time and the number in the system. Now, you see number; this is transform so; that means, the distribution is what you are talking about, the distribution of number $(P(z))$ in the system here $(F_T^*[\lambda(1-z)])$ the distribution of the system's waiting time. This is the PGF you are talking about, then what do you do? You take the Laplace-Stieltjes transform of the system waiting time which is T, and once you find that, then replace that variable by $\lambda(1-z)$, then that would be equal to $P(z)$. So this establishes then the relationship between the number in the system, the distribution of the number in the system, and the distribution of the sojourn time. So, this is the first relationship; we are going to use the relationship finally to put together everything together in a way; this is what you are observing first.

- By repeated differentiation of $P(z) = F_T^*[\lambda(1-z)]$, we can now find the relationship between moments of system size and system wait. By chain rule, we have that

$$\begin{aligned}\frac{d^k P(z)}{dz^k} &= (-1)^k \lambda^k \left. \frac{d^k F_T^*(u)}{du^k} \right|_{u=\lambda(1-z)} \\ &= (-1)^k \lambda^k (-1)^k \left. E\left[T^k e^{-Tu}\right] \right|_{u=\lambda(1-z)}\end{aligned}$$

  Now, we know PGF when you differentiate; when you put the variable tending to 1, you are going to get the factorial moments. So, on the left side, you will get the factorial moments.

- Let $L_(k)$ denote the Kth factorial moment of the system size and W_k the regular kth moment of the system waiting time. Then

$$\boxed{L_{(k)} = \left. \frac{d^k P(z)}{dz^k} \right|_{z=1} = \lambda^k W_k}$$

  So, this is what we are relating to now? We are relating the moments of these two distributions, in general, the kth order. But what we are getting now is that $L_{(k)}$ is the factorial moment, and $W_{(k)}$ is the simple moment or the regular moment of order k; $L_{(k)}$ is a factorial moment of order k. And so, this now, when $k = 1$, $L_{(1)}$ is simply the $E[N]$, and $W_{(1)}$ is $E[T]$, and this is simply λ. So, we are getting back the Little's law. Now, if it is 2, then this $L_{(k)}$ means what? It is, for example, if I write it for say $k = 2$ case. So, what will you get? On the left side, you will have $E[N[N-1]]$, and on the right side, what you will get, is $\lambda^2 E[T^2]$.

So, this is what you will get for $k = 2$ case; as an example, we are writing it out, but this is what basically this relationship tells you. So, this is the factorial moment on the left side; this is the ordinary moment on the right side. So, the moments when we talked about the question, whether what is the relationship between the higher order moments if there is one? Yes, there is a relationship, and the relationship is given by $\boxed{L_{(k)} = \left.\frac{d^k P(z)}{dz^k}\right|_{z=1} = \lambda^k W_k}$. So, this is then you can view it this is as a generalization of Little's law for higher-order moments. So now, suppose if I know this, then I can come to know $E[N^2]$ or $V[N]$ whatever, but essentially $E[N[N-1]] = \lambda^2 E[T^2]$ is the relationship that you are giving it here. So, in $k = 2$ case you will get $E[N[N-1]] = \lambda^2 E[T^2]$ this $k = 1$ you will get this Little's law and so on.

So, $\boxed{L_{(k)} = \left.\frac{d^k P(z)}{dz^k}\right|_{z=1} = \lambda^k W_k}$ is the generalization in terms of the moments, how the moments are related between the waiting time distribution and the number in the system in an $M/G/1$ system. So, this is the relationship we are establishing. So, this is the first question that we asked, and the answer to that is what is given by this generalization of Little's law. So, that is why we again put it in a boxed quantity. So, the first box quantity was $\boxed{P(z) = F_T^*[\lambda(1-z)]}$; the second one was $\boxed{L_{(k)} = \left.\frac{d^k P(z)}{dz^k}\right|_{z=1} = \lambda^k W_k}$ moment relationship.

Now, we will go; further; we look at the relationship between the distributions and what kind of relationship one can establish. Remember, we still have the distribution of the service time given in terms of the distribution in generic form, not a specific distribution that we have in mind. So, things have to boil down to that level; that is what is the relationship that we want to establish. Now, before we move, let us look at what happens in an $M/M/1$ queue.

- In $M/M/1$ queue, we looked at the system waiting time distribution can be written in terms of the service time distribution as this way

$$F_T(t) = (1-\rho)\sum_{n=0}^{\infty} \rho^n B^{(n+1)}(t),$$

  where $B^{(n+1)}(t)$ is the $(n+1)-$ fold convolution of the exponential CDF $B(t)$.

  ▶ Memoryless property helped to get the above result, taking care of the situation that the arrivals catch the server in the middle of a serving period with probability ρ.

  ▶ We do not have memoryless property now (in general unless the distribution itself is exponential), and hence we require an alternative approach to derive a comparable result for $M/G/1$.

So, we need an alternative way of deriving such a comparable result for an $M/G/1$ queue; that is what we will aim to do. So, towards that end, what we will do?

- We start with deriving a simple relationship between the LSTs of the service and waiting times (both of them are continuous in a way), $B^*(s)$ and $F_T^*(s)$. We know that $P(z) = F_T^*[\lambda(1-z)]$ and, from our earlier results, we also know that

$$P(z) = \mathit{\Pi}(z) = \frac{(1-\rho)(1-z)K(z)}{K(z-z)}$$

But the PGF $K(z)$ of $\{k_i\}$ (the number of arrivals during the service time) is given by

$$K(z) = \int_0^\infty e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda t z)^n}{n!} dB(t) = \int_0^\infty e^{-\lambda t(1-z)} dB(t)$$

That is $\boxed{K(z) = B^*[\lambda(1-z)]}$. So, this is what we are observing now.

So, this is another fact that we are observing. So, we have already observed $\boxed{P(z) = F_T^*[\lambda(1-z)]}$ fact, and we know about $P(z) = \Pi(z) = \frac{(1-\rho)(1-z)K(z)}{K(z-z)}$. Now $\boxed{K(z) = B^*[\lambda(1-z)]}$ here.

- We will now put all of these together to see what happens.
- Putting all these three facts together, we obtain

$$F_T^*[\lambda(1-z)] = \frac{(1-\rho)(1-z)B^{(*)}[\lambda(1-z)]}{B^*[\lambda(1-z)] - z}$$

$$\boxed{F_T^*(s) = \frac{(1-\rho)sB^*(s)}{s - \lambda[1 - B^*(s)]}}$$

You will arrive at this expression which relates the system time to the service time ultimately, that is what we wanted; if you want to get the distribution of the system waiting time, the service time distribution should be known. Now, once service time distribution is known, that is, in terms of its transforms, I can substitute here and get this distribution. And if I invert that, if I take the inversion of that, I will be going to get the waiting time distribution like in the case of. So, this is what you see here; for the system waiting time or the sojourn time or waiting time in the system.

- Now, since $T = T_q + S$, we know, from the convolution property of transforms, that $F_T^*(s) = F_T^*(s)B^*(s)$. Thus,

$$\boxed{F_T^*(s) = \frac{(1-\rho)s}{s - \lambda[1 - B^*(s)]}}$$

So, at least now, in terms of the transforms of the service time distribution, you have given now the system waiting time. And the sojourn time or waiting time in the system $F_T^*(s)$ and waiting time in the queue $(F_{T_q}^*(s))$, now once I know what this $B^*(s)$ is, I get this expression. Now, suppose if I have an exponential distribution, then I know what is this $B^*(s)$ just substitute it, and you will get the distribution. You should get that same distribution that you have derived earlier. But now, you can go much beyond exponential to see what distributions for which you have. So, given; this now, I can obtain the distribution, but now the distribution we are writing it in terms of the transforms, that is what you see here. Now, $F_T^*(s) = \dfrac{(1-\rho)s}{s - \lambda[1 - B^*(s)]}$, say, suppose now if you want a little bit more explicitly because we want to see a similar result to what happens in the $M/M/1$ case.

- Now, expanding the right hand side as a geometric series (since $(\lambda/s)[1 - B^*(s)] < 1$),

$$F_{T_q}^*(s) = (1-\rho)\sum_{n=0}^{\infty}\left(\frac{\lambda}{s}[1 - B^*(s)]\right)^n = (1-\rho)\sum_{n=0}^{\infty}\rho^n\left(\frac{\mu}{s}[1 - B^*(s)]\right)^n.$$

- It can be seen that $\mu[1 - B^*(s)]/s$ is the LST of the residual service time distribution

$$R(t) = \mu \int_0^t [1 - B(x)]dx.$$

Recall our renewal theory results, but there rate we assumed it as $1/\mu$, but here the rate is μ. So, this term appears here as μ, but in renewal theory, we call $1/\mu$ as the rate. So, that was $1/\mu$ here, exactly the same expression that we are getting it here.

► $R(t)$ is the CDF of the remaining service time of the customer being served at the time of an arbitrary arrival, given that, of course, the arrival occurs when the server is busy.

- We thus obtain

$$F^*_{T_q}(s) = (1 - \rho) \sum_{n=0}^{\infty} \rho^n [R^*(s)]^n$$

.

Now, once I have this form if I can take the inversion of this process or the Laplace-Stieltjes transform, then because I know that if I have the power then in the inversion, I will get the convolution.

After term-by-term inversion utilizing the convolution property, we obtain

$$F_{T_q}(t) = (1 - \rho) \sum_{n=0}^{\infty} \rho^n R^{(n)}(t)$$

So, I will end up with this expression which is a result, which is similar in look to the case of an $M/M/1$, but it does not involve the same thing. But it is something similar one can give an argument; one can give an interpretation to such kind of expression, what it means, and so on.

So, that is what then one can do. So, basically like, if you reorder the time with the remaining service time as the fundamental unit, which is what this $R(t)$. Then the steady states find n such time units of potential service in front of it, with this probability $(1 - \rho)^n$ which is if then if you interpret in that form, then it is similar to what one would give it for, in the case of an $M/M/1$ model. So, that is what we are seeing it here. But mainly like this is an additional input, but mainly this part is what is relevant, that if you are looking for the distribution of service system waiting time or queue waiting time, then they are related to the service time distribution in this form. Now, once I know what $B^*(s)$ is, then I can substitute here, and then you can do it. So, as an **exercise**, take exponential and derive result, similar to or which will result in $M/M/1$ result, not similar in a way which is exactly will be equal to the $M/M/1$ result.

So, you can do that exercise to get the idea; that is what waiting time distributions that we have been talking about. Now, similarly, one can also conduct a busy period analysis, but here the busy period analysis is a bit in terms of the time domain you are looking at it; it is a bit more difficult than what one would do for $M/M/1$ and $M/M/1$ itself we know that how difficult it was.

But it is not that difficult to find the Laplace-Stieltjes transform of the busy period and from which at least the moments can be obtained. So, one can also get that easily. How do we do? Let us proceed because a busy period is also an important concept in a queue in general, we have done it only for $M/M/1$, but for $M/G/1$, we can give it so that you can at least get the moments of this busy period which will be very relevant in many practical situations.

- Let $G(x)$ denote the CDF of the busy period ; (you already know what a busy period means. There is an empty system to which a first customer comes, and from that point till when the server becomes idle for the first time. So, this is what is called the busy period).
- So, now we are looking at the duration or the time duration, and its distribution is what we are looking at it here. Now, let us take an $M/G/1$, with the service CDF $B(t)$, let $G(x)$ denote the CDF of the busy period, which we call some X some random variable. Now, what can we do?

► We can condition on this X on the length of the first service time inaugurating the busy period.
So, we can condition on this particular X, which is the length of the first duration first service time because it is the empty system to which the first customer comes. Now, during his service time, how many would have come and each one will generate its own busy period, because during its; suppose, there are 5 customers has come to assume. Now, the first customer when during his service time, also, some number would have come. So, that will be probabilistically equal to the very first customer that you consider; during his service time, these 5 customers have come. Now, out of these 5, the first one will be the next to go to the first service. Now, during his service time again, some more would have come, that is equivalent to original customer that you can call it. So, each arrival during the first service time of the busy period generates its own busy period. So, you have to keep that in mind to write this one.

Now, if I want $G(x)$, which means the total busy period is less than or equal to x. So, if I want X, then $G(x) = Pr\{X \leq x\}$; that is what this $G(x)$ is.

$$\begin{aligned} G(x) &= \int_0^x Pr\{\text{busy period generated by all arrivals during} t \leq x - t | \text{first service time =} t\} dB(t) \\ &= \int_0^x \sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!} G^{(n)}(x-t) dB(t) \end{aligned}$$

where $G^{(n)}(x)$ is the $n-$fold convolution of $G(x)$.
Next, let $G^*(s)$ be the LST of $G(x)$, and $B^*(s)$ be the LST of $B(t)$.

- Then, by taking the Laplace-Stieltjes transform both sides of the above, it is found that

$$G^*(s) = \int_0^{\infty} \int_0^x \sum_{n=0}^{\infty} \frac{e^{-sx}(\lambda t)^n}{n!} G^{(n)}(x-t) dB(t) dx.$$

- Changing the order of integration

$$G^*(s) = \int_0^\infty \int_t^\infty \sum_{n=0}^\infty e^{-sx} \frac{e^{-\lambda t}(\lambda t)^n}{n!} G^{(n)}(x-t)dB(t)dx.$$
$$= \int_0^\infty \sum_{n=0}^\infty \frac{e^{\lambda t}(\lambda t)^n}{n!} \int_t^\infty e^{-sx} G^{(n)}(x-t)dxdB(t).$$

- Applying a change of variable $y = x - t$ to the inner integral gives

$$G^*(s) = \int_0^\infty \sum_{n=0}^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!} \left(e^{-st} \int_0^\infty e^{-sy} G^{(n)}(y)dy \right) dB(t)$$

- By the convolution property

$$G^*(s) = \int_0^\infty \sum_{n=0}^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!} e^{-st}[G^*(s)]^n dB(t) = \int_0^\infty e^{-\lambda t} e^{\lambda t G^*(s)} e^{-st} dB(t)$$

That is $\boxed{G^*(s) = B^*[s + \lambda - \lambda G^*(s)]}$.

It is not in explicit form like the other cases, like in the system waiting time, what we did is T or T_q we expressed in terms of B completely, but here B you need, but within B again there is a G^*, it is an interlinked relationship what then you are seeing it. Of course, you need the form of this B star. So, once you have this Laplace-Stieltjes transform of the service time distribution, then the argument you have to substitute $s + \lambda - \lambda G^*(s)$ which will again involve $G^*(s)$, which is what you are trying to compute. So, it is something like an integral equation kind of thing; essentially, that is what it is. So, that is what you are getting it here.

So, this is the relationship that is the main part that how the distribution of the busy period is related to the distribution of the service time because everything else is specified. Now, the service time is what is unspecified, which you will check because this is the $M/G/1$ model. So, that relationship is $\boxed{G^*(s) = B^*[s + \lambda - \lambda G^*(s)]}$. Now, depending upon this B^*, one can do some solution of this whether we can obtain and so on one can think. Maybe again, exponential case you can think, deterministic case you can think, and so on, some simpler cases, you can always work it out to see like how one can obtain the busy period in that case, but nevertheless, from this equation itself, one can obtain the mean values.

Say, for example, if you are looking at the mean length of the busy period, how can one obtain it?

$$E[X] = -\left. \frac{dG^*(s)}{ds} \right|_{s=0} \equiv G^{*\prime}(0),$$

where $G^{*\prime}(s) = B^{*\prime}[s + \lambda - \lambda G^*(s)] \left[1 - \lambda G^{*\prime}(s)\right]$

- Therefore,

$$E[X] = -B^{*\prime}[\lambda - \lambda G^*(0)] \left[1 - \lambda G^{*\prime}(0)\right] = -B^{*\prime}(0)\{1 + \lambda E[X]\},$$

or

$$E[X] = -\frac{B^{*\prime}(0)}{1 + \lambda B^{*\prime}(0)}$$

Because $B^{*\prime}(0) = -\frac{1}{\mu}$,

$$E[X] = \frac{1/\mu}{1 - \lambda/\mu} = \frac{1}{\mu - \lambda}$$

- This is exactly the same result we obtained earlier for $M/M/1$.

And at that time, we said that this is also true for $M/G/1$ case. And here is what you are seeing in general for this particular case because this is mean we are assuming we do not mean that the B is exponential here. We are only looking at the mean of the service time distribution, which we are denoting in terms of μ. So, it is $1/\mu$ is what is the mean. So, μ is the rate that you are talking about, so this is the exactly same result. So, this also means, in some sense, the insensitivity to the service time distribution you have; as long as the mean is known, you will get $E[X] = \frac{1/\mu}{1-\lambda/\mu} = \frac{1}{\mu-\lambda}$ as the mean busy period that you have seen here. So, this is what we said in the prototype; now, you can see that this is also true for $M/G/1$.

And $M/G/1$ because in generality like busy period since it is an important component you need to understand. So, basically, the two main things here is $\boxed{G^*(s) = B^*[s + \lambda - \lambda G^*(s)]}$. relationship between the Laplace transform of the busy period CDF and the CDF of the service time distribution is related by this relationship, and $E[X] = \dfrac{1/\mu}{1 - \lambda/\mu} = \dfrac{1}{\mu - \lambda}$ gives the mean. Now, higher order moments again, one can obtain using the same relationship that we are here.

So, this is what we have it for $M/G/1$ queue; I mean, with infinite capacity and FCFS discipline. So, we have seen again, like in the $M/M/1$ case, here also we have characterized the number in the system, the waiting times, the busy period everything we have done for $M/G/1$. Now, like when G is equal to an exponential distribution, then you will get the corresponding results of $M/M/1$ very nicely. But now, you can have a much more generic distribution also, but now because of the assumptions that you make, which is a general time distribution, what we are seeing is that the results are all mainly in transform form. But from which moments can be obtained. Now, once you have this specific form of the service time distribution, then you can get the transform results for the number in the system or service time, which inversion, whether it is z transform inversion or the Laplace-Stieltjes transform inversion, you will get the actual distribution is what can give. So, this is the broad picture, and that is what we have done for $M/G/1$ model. So, let us stop here; we will continue with something more in the later classes.
Thank you, bye.