Introduction to Queueing Theory

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

Lecture - 44
G/G/1 Queues: Lindley's Integral Equation

Hi and hello, everyone. What we have seen so far are either Markovian queueing systems or semi Markovian queueing systems, where at least one of the arrival process or service processes we have assumed to be following the Markovian structure, meaning that exponential distribution was there. Even when we generalize to general Markovian or one can do in a similar way to general semi Markovian, we still can retain some amount of Markovian structure by restricting to either inter-arrival distribution or service time distribution to either exponential or something that can be expressed in terms of exponential so that we can handle it in a way. But what we will see in this final phase is that we do not have any of those. So, what we have? We have what we generally refer to as general queueing models. We are not calling it as Markovian or semi Markovian or anything of that sort; it is a general queueing model. So, what we have? We have an input process, a service process, and the number of servers we will restrict to single only; otherwise, things will become more complex. So, single, but the general input meaning that it is no more Poisson or interarrival time or no more Poisson or Erlang or phase-type or anything we are not assuming, it is general, it could be any distribution. Similarly, the service time distribution could be any distribution. In this kind of generality, when you look at it, obviously, it is not going to be that easy to analyze unless you put some such restrictions. So, what researchers have done is that even within this, they try to have some sort of restriction on the kind of interarrival times that you can have or the service time that you can have so that you can get something out of it, that is always there. But and that is a way one should look at as well. But we will try to see in such generality itself how one can handle and how difficult it is, and the reason why then people want to look at with some restrictions would be clear to you in that sense. So, what we consider is what we call it as $G/G/1$ queues. So, it is general input, general service, single server queueing system. We still have these interarrival times are IID, same as the service times are also IID, and they are mutually independent; I mean, interarrival times and service times they are all independent of each other.

The interarrival times the distribution is we are giving it in terms of this A, which is the CDF of the interarrival time; they are IID, so it is all of them have the same A. Similarly, service times have the CDF as B. As we said, though, because of, in say, generality, we do not expect any specific structure for the model to be there; one can still obtain some results. What we will obtain it is what is called as an integral equation, which is basically called as a Wiener-Hopf type integral equation for the stationary distribution of the waiting time in a queue of an arbitrary customer, and this is also called Lindley's equation or Lindley's integral equation. This work is called Lindley's equation in queueing theory mainly because of the contribution; all of this was contributed by Lindley in the 1950s. So this that is why it is called as Lindley's equation, and this also another approach; of course, this approach you can use it now for your semi Markovian model or Markovian model; anything you can also use this approach should result in the same expression that you expect to have. So, that is why this Lindley's integral equation approach is also one of the approaches to

handle, say, semi Markovian systems; apart from embedded Markov chain supplementary variable, this is also another technique that one can say. For more details on this $G/G/1$, one can look at the single server queue book by Cohen (1982); but many of those results are beyond our scope. So, we will restrict to something we can handle; but in this lecture, what we are going to see is that things are a little complex, obviously.

This is just to show you how complex things can become if you want to handle it in such generality, and in the later lecture, we will try to see like what we can obtain in an easier way. So, in this case, we will just try to exhibit that, but we just want to see how one can analyze that. So, that is what we want to see; but this requires the idea of how one can solve this Wiener-Hopf kind of integral equation and so on, which you may or may not be knowing but let us not worry about it; you do not need to think too much on that. So nevertheless, you can just see like what is going on here; that is what is more important. But if required, obviously, you will go deep into that to understand and if you have to analyze it; but at least you should have a feel of how things can get complex in such a situation and how one can handle it, that is what is the idea of this particular lecture. So, this is the background that we have here. Now, what we are trying to do is that this Lindley's equation is what we will try to obtain, which is an integral equation of Wiener-Hopf type for the stationary distribution of the waiting time in queue of an arbitrary customer. So, let us get started.

- So, we start with the relationship between the line waiting time, the waiting times in the queue, or line waiting time or line delay; it is what is the common word that we use, which we call $W_q^{(n)}$ and W_q^{n+1}.

  So, we want to frame or formulate a relationship between these two, which is of the nth customer what is his line delay and $(n+1)$th customer what is will be his or her line delay. So, now this is valid for any arbitrary $G/G/1$ problem. What is the relationship? The relationship is if we want to look at what would be the line delay or the delay in queue or waiting time in the queue for this particular arbitrary customer of $(n+1)$th customer is would be equal to this. What is this?

$$W_q^{(n+1)} = \begin{cases} W_q^{(n)} + S^{(n)} - T^{(n)}, & W_q^{(n)} + S^{(n)} - T^{(n)} > 0, \\ 0, & W_q^{(n)} + S^{(n)} - T^{(n)} \leq 0 \end{cases}$$

  You can see; suppose if nth customer has to wait for, say 5 time units; now what would be his waiting time, $(n+1)$th waiting time? Of course, this went for service for 3 units. So, $5+3=8$, and the interarrival time; means when the next customer arrived; if it is after 2 units, so then his total waiting time would be 6 units. So, it is obvious you can see, and this will be true only if $W_q^{(n)} + S^{(n)} - T^{(n)} > 0$. If $W_q^{(n)} + S^{(n)} - T^{(n)} \leq 0$, obviously, his wait in the delay is 0, which means that system has become ideal already; if this quantity is less than or equal to 0. So, that is the relationship that you can formulate between the line waiting times of the nth customer and the $n+1$th customer in terms of the service time of the nth customer and the time between the arrival of nth customer and $(n+1)$th customer, which is T_n, which together you can write

$$W_q^{(n+1)} = \max(0, W_q^{(n)} + S^{(n)} - T^{(n)})$$

  Here $S^{(n)}$ is the service time of the n^{th} customer and $T^{(n)}$ is the time between the arrivals of the two customers.

- The stochastic process $\{W_q^{(n)}, n = 0, 1, 2, \dots\}$ is a discrete-time Markov process, since the behaviour of $W_q^{(n+1)}$ is only a function of the stochastically determined value of $W_q^{(n)}$ and is independent of prior waiting-time history.

- From basic probability arguments

$$\begin{aligned} W_q^{(n+1)}(t) &= P\{\text{line delay } W_q^{(n+1)} \text{of } (n+1)st \text{ customer} \leq t\} \\ &= P\{W_q^{(n+1)} = 0\} + P\{0 < W_q^{(n+1)} \leq t\} \\ &= P\{W_q^{(n)} + S^{(n)} - T^{(n)} \leq 0\} + P\{0 < W_q^{(n)} + S^{(n)} - T^{(n)} \leq t\} \\ &= P\{W_q^{(n)} + S^{(n)} - T^{(n)} \leq t\}. \end{aligned}$$

- Define $U^{(n)} = S^{(n)} - T^{(n)}$ and let $U^{(n)}(x)$ denotes the CDF. Then, we have by the convolution formula

$$W_q^{(n+1)}(t) = \int_{-\infty}^{t} W_q^{(n)}(t-x) dU^{(n)}(x) \quad (0 \leq t < \infty).$$

Now, in the steady-state, we assume that when we assume these interarrival times and service times, we generally assume that it has a finite mean and finite variance, that is what normally you encounter in practice. So, that is, we assume, and the mean for interarrival time is always you assume to be $1/\lambda$, and for the service time, the mean is assumed to be $1/\mu$. So, this ρ is basically still λ/μ, and if you for steady-state, we need the stability condition as $\rho < 1$. And under that situation, the two waiting time CDFs must also be identical, which means that $W_q^{(n)}$ and $W^{(n+1)}$, as you take $n \to \infty$, and we denote the CDF of the steady-state waiting time in queue distribution, to be $W_q(t)$. Remember earlier we the random variables, that random variable we call it as T_q to denote the line delay and F_{T_q} as the CDF of that random variable. But we are slightly changing the notation for this $G/G/1$ model alone, which is basically we are denoting it by this quantity itself. So, that is what we write, observe the difference in notation over what we have been using so far with respect to this particular random variable and the corresponding CDF. Now, if you take $\lim n \to \infty$; then what we have

$$W_q(t) = \begin{cases} \int_{-\infty}^{t} W_q(t-x) dU(x) & (0 \leq t < \infty) \\ 0 & (t < 0) \end{cases}$$

This is what is called as Lindley's equation; of course, you can write it in different forms. So, the same thing you can also write

$$W_q(t) = -\int_{0}^{\infty} W_q(y) dU(t-y) \quad (0 \leq t < \infty)$$

there is one another form, or other forms also you can write it, whatever is convenient to you one can work on it. So, $W_q(t) = \begin{cases} \int_{-\infty}^{t} W_q(t-x) dU(x) & (0 \leq t < \infty) \\ 0 & (t < 0) \end{cases}$ is what is called as Lindley's integral equation because this is an integral equation because you want to determine the quantity $W_q(t)$. But $W_q(t)$ is not only appearing here but is also as part of an integrand, within this integral, whereas this $U(x)$ is given in a way. So, this is what is that integral equation. And this can also be written in a modified form $W_q(t) = -\int_0^{\infty} W_q(y) dU(t-y) \quad (0 \leq t < \infty)$; there are other modified forms also; one can write where this $U(x)$ is again the equilibrium version of this $U^{(n)}(x)$, which is basically the convolution of S and $-T$, which can basically be written in this form.

$$U(x) = \int_{\max(0,x)}^{\infty} B(y)dA(y-x).$$

So, this is if I can solve this integral equation; say in mathematics there are in many areas that problems, the solutions for which can be obtained as in a functional form as an integral equation and solving that integral equation we can obtain the explicit solution. So, here also, we are looking at the line delay distribution, which is given in terms of this Lindley's integral equation or simply Lindley's equation, which is a Wiener-Hopf type integral equation. So, a solution to this will give me the distribution of the line delay in this $G/G/1$ model. And once I have this distribution of line delay, I can obtain the average, and by Little's law and other stuff, then I can obtain the other three quantities as well. So, the basic four quantities one can obtain once I have this. So, that was the idea. So, this $U(x)$ is basically the convolution of S and $-T$; because this is U, the random variable is basically $S - T$.

So, thus what is our conclusion at this stage is that the distribution of the line delay depends only on the distribution of the difference between the service time and the inter-arrival distribution, rather than on the individual distributions.

Thus, we see that the distribution of line delay depends only on the distribution of the difference between the service time and interarrival time distributions, rather than on the individual distributions. Because this W_q involves U, not A and B individually, but this $U(x) = \int_{\max(0,x)}^{\infty} B(y)dA(y-x)$. Now, like one can have the same U for different B and A possibly. So, we are not worried about what are this individual distribution; what is that we want is this U. Once we have this U, this distribution; then this equation can be solved to get the CDF of the line delay; that is what we see ultimately out of this Lindley's integral equation. So, this

$$W_q(t) = \begin{cases} \int_{-\infty}^{t} W_q(t-x)dU(x) & (0 \le t < \infty) \\ 0 & (t < 0) \end{cases}$$
$$= -\int_{0}^{\infty} W_q(y)dU(t-y) \quad (0 \le t < \infty)$$

is the main quantity which we call as Lindley's equation. So, that is what we see that line delay depends on the difference between $S - T$; basically, it depends on U, U is basically $S - T$, rather than individually on S, and T is what we are seeing it here. So, one has to solve this to get the distribution, the Lindley distribution. So, let us see how one can solve that. If this looks like a convolution, but it is not exactly so, and it is Wiener-Hopf integral equation, now how one can solve this.

- To solve such an equation, what we do is we define a new function, which is also sometimes called complementary waiting time because, from the look of $W_q(t)$, this is defined for $t \ge 0$. Now, in the reverse way, for $t < 0$, you can define it in this way

$$W_q^-(t) = \begin{cases} \int_{-\infty}^{t} W_q(t-x)dU(x) & (t < 0), \\ 0 & (t \ge 0), \end{cases}$$

  Then from Lindley's equation

$$W_q^-(t) + W_q(t) = \int_{-\infty}^{t} W_q(t-x)dU(x) \quad (-\infty < t < \infty). \tag{1}$$

Note that $W_q^-(t)$ is the portion of the CDF associated with the negative values of $W_q^{(n)} + S - T$ when there is idle time between the n^{th} and the $(n+1)^{st}$ customer.

Because our distribution is on the non-negative side, we put the whole mass at 0, and we call that is some quantity. Even you have seen in $M/M/1$; there is a mass at time 0. Now, if you distribute it to our the negative real axis, what you would get is what is basically you will get from

$W_q^-(t) = \begin{cases} \int_{-\infty}^{t} W_q(t-x)dU(x) & (t<0), \\ 0 & (t \geq 0), \end{cases}$. This is that function; one can relate to that quantity. So, $W_q^-(t) + W_q(t) = \int_{-\infty}^{t} W_q(t-x)dU(x) \quad (-\infty < t < \infty)$. is what you will obtain after defining this.

- It turns out to be easiest to try to obtain $W_q(t)$ for $t > 0$, since $W_q(t)$ is not continuous at 0, but has a jump equal to the arrival-point probability a_0, so that $W_q(0) = a_0$.

  Denote the two-sided Laplace transforms of $W_q(t)$ and $W_q^-(t)$ as (do not worry too much one what these are, how it is different, and so on; it is just that in Laplace transform, you see that it is defined from 0 to infinity. So, this is and the both sides you are defining it, you can think of it as if this $\int_{-\infty}^{\infty} e^{-st} W_q(t)dt$, rather than 0 to t. If you want more to know about it, you can explore no problem; but that is what it is)

$$\overline{W}_q(s) = \int_{-\infty}^{\infty} e^{-st} W_q(t)dt = \int_{0}^{\infty} e^{-st} W_q(t)dt$$
$$\overline{W}_q^-(s) = \int_{-\infty}^{\infty} e^{-st} W_q^-(t)dt = \int_{-\infty}^{0} e^{-st} W_q^-(t)dt.$$

  We shall also use $U^*(s)$ as the (two-sided) LST of $U(t)$.

- We take the two-sided Laplace transform of both the sides of $W_q^-(t) + W_q(t) = \int_{-\infty}^{t} W_q(t-x)dU(x) \quad (-\infty < t < \infty)$.. The transform of the right-hand side is

$$\mathcal{L}_2 \left\{ \int_{-\infty}^{t} W_q(t-x)dU(x) \right\} = \int_{-\infty}^{\infty} \int_{-\infty}^{t} e^{-(t-x)s} W_q(t-x) e^{-sx} dU(x)dt.$$

  Since $W_q(t-x) = 0$ for $x \geq t$, we can write

$$\begin{aligned} \mathcal{L}_2 \left\{ \int_{-\infty}^{t} W_q(t-x)dU(x) \right\} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(t-x)s} W_q(t-x) e^{-sx} dU(x)dt. \\ &= \left(\int_{-\infty}^{\infty} e^{-su} W_q(u)du \right) \left(\int_{-\infty}^{\infty} e^{-sx} dU(x) \right) \\ &= \overline{W}_q(s) U^*(s). \end{aligned}$$

  But U is the CDF of the difference of the interarrival and service times and hence by the convolution property must have (two-sided) LST equal to the product of the interarrival transform $A^*(s)$ evaluated at $-s$ and the service transform $B^*(s)$, since $A(t)$ and $B(t)$ are both zero for $t < 0$.

  Hence $U^*(s) = A^*(-s)B^*(s)$, and from (1),

$$\overline{W}_q^-(s) + \overline{W}_q(s) = \overline{W}_q(s) A^*(-s) B^*(s)$$
$$\implies \overline{W}_q(s) = \frac{\overline{W}_q^-(s)}{A^*(-s)B^*(s) - 1}$$

- Therefore, given any pair $(A(t), B(t))$ for the $G/G/1$, we can theoretically find the Laplace transform of the line delay.
- The determination of $\overline{W}_q^-(s)$ is the primary difficulty in this computation, often requiring advanced concepts from the theory of complex variables.
- But there are some specialized procedures, say, for example, a spectral method or some such thing.

If you assume more on something on the inter-arrival and service time distribution. Say, for example, if you assume that the transforms have this rational form for this arrival and service time distribution, then one can adopt a slightly easier procedure to get $\overline{W}_q^-(s)$ and hence the line delay transform. And by inversion of that, you can obtain the CDF of the line-delay random variable. So, this is what is the essence now given; but again, solving Lindley's equation is the process. Now, $\overline{W}_q^-(s)$ requires that you define a new function, and its determination is what is the main bottleneck in the determination of the line delay, but now you are explicitly used A and B in this form. So, this is another major quantity that that is why we might put it in the box.

$$\boxed{\overline{W}_q(s) = \frac{\overline{W}_q^-(s)}{A^*(-s)B^*(s) - 1}}$$

So, this is what is most relevant; once I know given A and B, if I know this, then I know this, that is what it is. So, I determine $\overline{W}_q^-(s)$.

So, in general, it is difficult. So, there is a difficulty, main difficulty is determination of $\overline{W}_q^-(s)$. Now, to understand to show how that can be determined or how it can be worked out, we will take up the very simplest case of this $M/M/1$ model. So, from $G/G/1$, now we will go to $M/M/1$, so just to show the working as to how one can determine that particular quantity, which is $\overline{W}_q^-(s)$.

- To show the working, let us consider the $M/M/1$ problem for which

$$B(t) = 1 - e^{-\mu t}, \quad B^*(s) = \frac{\mu}{\mu + s}, \quad A(t) = 1 - e^{-\lambda t} \text{ and } A^*(-s) = \frac{\lambda}{\lambda - s}$$

- Using the equation $U(x) = \int_{\max(0,x)}^{\infty} B(y)dA(y - x)$, we have

$$U(x) = \begin{cases} \int_0^\infty (1 - e^{-\mu y})\lambda e^{-\lambda(y-x)}dy & (x < 0), \\ \int_x^\infty (1 - e^{-\mu y})\lambda e^{-\lambda(y-x)}dy & (x \geq 0) \end{cases}$$

$$= \begin{cases} \dfrac{\mu e^{\lambda x}}{\lambda + \mu} & (x < 0) \\ 1 - \dfrac{\lambda e^{-\mu x}}{\lambda + \mu} & (x \geq 0). \end{cases}$$

Thus

$$W_q^-(t) = \int_{-\infty}^{t} W_q(t - x)dU(x) \quad (t < 0)$$

$$= \frac{\lambda\mu}{\lambda + \mu} \int_{-\infty}^{t} W_q(t - x)e^{\lambda x}dx \quad (t < 0)$$

Letting $u = t - x$ yields

$$\begin{aligned} W_q^-(t) &= \frac{\lambda\mu}{\lambda+\mu}\int_0^\infty W_q(u)e^{-\lambda(u-t)}du \\ &= \frac{\lambda\mu e^{\lambda t}}{\lambda+\mu}\int_0^\infty W_q(u)e^{-\lambda u}du \\ &= \frac{\lambda\mu e^{\lambda t}\overline{W}_q(\lambda)}{\lambda+\mu} \end{aligned}$$

Now we have to find $\overline{W}_q(\lambda)$.

- For any $M/G/c$ queue, we have that

$$\pi_n^q = P\{n \text{ in queue just after a departure}\} = \frac{1}{n!}\int_0^\infty (\lambda t)^n e^{-\lambda t} dW_q(t)$$

  Hence, if $G = M$ and $c = 1$, we find that

$$\begin{aligned} \pi_0^q &= \int_0^\infty e^{-\lambda t} dW_q(t) \\ &= e^{-\lambda t}W_q(t)|_0^\infty + \lambda\int_0^\infty e^{-\lambda t}W_q(t)dt. \end{aligned}$$

- But $\lim_{t\to\infty} e^{-\lambda t} = 0$. Also, since we are only concerned computationally with $W_q(t)$ for $t > 0$, let us make $W_q(0) = 0$ to simplify the analysis in the sequel. In the end, we will simply set $W_q(0) = p_0$, since it is true for all $M/G/1$ that $W_q(0) = p_0 = 1 - \lambda/\mu$.

- Therefore, we have

$$\pi_0^q = \lambda\overline{W}_q(\lambda).$$

- Also we have $\pi_0^q = p_0 + p_1 = (1-\rho)(1+\rho)$ (as we are doing the analysis for $M/M/1$). Hence

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

- By using the equation $W_q^-(t) = \dfrac{\lambda\mu e^{\lambda t}\overline{W}_q(\lambda)}{\lambda+\mu}$, we get

$$W_q^-(t) = \frac{e^{\lambda t}(1-\rho)(1+\rho)}{1+\rho} = e^{\lambda t}(1-\rho)$$

  with transform

$$\overline{W}_q^-(s) = \frac{1-\rho}{\lambda - s}$$

- Now put everything together and using equation $\overline{W}_q(s) = \dfrac{\overline{W}_q^-(s)}{A^*(-s)B^*(s) - 1}$, we find that

$$\begin{aligned} \overline{W}_q(s) &= \frac{(1-\rho)/(\lambda-s)}{\lambda\mu/[(\lambda-s)(\mu+s)] - 1} = \frac{(1-\rho)(\mu+s)}{s(\mu-\lambda+s)} \\ &= \frac{1-\rho}{s} + \frac{\lambda(1-\rho)}{s(\mu-\lambda+s)}. \end{aligned}$$

which on inversion yields

$$W_q(t) = 1 - \rho + \frac{\lambda(1-\rho)(1-e^{-(\mu-\lambda)t})}{\mu-\lambda}$$
$$= 1 - \rho e^{-\mu(1-\rho)t} \quad (t > 0).$$

Now, realizing that $W_q(0)$ equals $p_0 = 1 - \rho$, the result is the same as obtained for $M/M/1$.

So, this is how you work when you have this kind of $G/G/1$; even for $M/M/1$, things are complex, but even if you have to use other general distribution, things are going to be much more complex. But you need more ideas from complex analysis theory to determine $\overline{W}_q^-(s)$; whatever this quantity, that is what is the main point here. So, what is a summary here? In general, analyzing $G/G/1$ queues involves complex analysis as well as transform inversion processes; because you want to get in the time domain, so you need to get. If you want to mean, of course, directly, you will get that. Generally involves these two, and these two procedures are complicated and are very much involved, it is not that easy to see. So, in many cases, we need numerical methods to perform the transform inversion. So, there are approaches available to numerically invert the transforms. But Laplace transform inversion is not an easy thing to do, unlike the Fourier transform inversion. So, there are issues involved with that. So, that is why it also involves a lot of effort on that side. But what we can do is that rather than getting this kind of computation distribution and hence the mean, one can maybe try to get some kind of bounds and approximations for the analysis of such $G/G/1$; we are not going to take it up this approximation in our course. But we will talk about a little bit about bounds for the performance measures with respect to $G/G/1$ and how that can be easily obtained, which can give us a lot of ideas about the system under consideration, which many cases might be sufficient enough to handle at least in the broader picture. So, that is what we might take it up next. So, as you can see, the complete analysis involves a lot of effort from, and you require more ideas from different fields of mathematics to do that kind of analysis. That is what we wanted to highlight in this, and that probably like you would have got an idea of this.
Thank you, bye.